FEIGNING COGNITIVE DEFICITS ON NEUROPSYCHOLOGICAL EVALUATIONS: MULTIPLE DETECTION STRATEGIES

Scott D. Bender, B. A., M. S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2000

APPROVED:

Richard Rogers, Major Professor
James Quinn, Committee Member
Kenneth Sewell, Committee Member
Ernest Harrell, Committee Member and Chair of the Department of Psychology
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

Bender, Scott D., Feigning cognitive deficits on neuropsychological evaluations: Multiple detection strategies. Doctor of Philosophy (Clinical Psychology), December 2000, 190 pp., 43 tables, 3 figures, 127 references.

Individuals undergoing forensic neuropsychological evaluation frequently stand to gain in some manner if neurocognitive dysfunction is diagnosed. As a result, neuropsychologists are customarily asked to test for neurocognitive feigning during the assessment. The current study employed an analogue design with a clinical comparison group to examine the utility of the TOCA (Rogers, 1996) as a measure of feigned neurocognitive impairment. Two groups of simulators (one cautioned about the presence of detection strategies and one not cautioned) were compared to clinical and normal control groups. Fourteen scales were developed based on five detection strategies: symptom validity testing, performance curve, magnitude of error, response time, and floor effect. Each was employed during both verbal and nonverbal tasks. Significant differences were revealed among groups when subjected to ANOVA. Classification rates from subsequent utility estimates and discriminant function analyses on the scales ranged from 58.8% to 100%. Combining strategies yielded a classification rate of 95.7%. The effect of cautioning simulators was modest; however, a trend was noted on some scales for cautioned simulators to appear less obviously impaired than noncautioned. Although the results require crossvalidation, preliminary data suggest that the TOCA is a sensitive and specific measure of feigned neurocognitive performance. Strengths and weaknesses of the study are discussed and directions for future research are proposed.

ACKNOWLEDGEMENTS

I would like to thank the members of my dissertation committee who contributed a great deal of time and expertise to this project: Drs. Richard Rogers, Kenneth Sewell, Ernest Harrell, and James Quinn. I would also like to thank Drs. Duhon and Houtz for their assistance in enlisting participants and for allowing me to use their offices for data collection. Thanks to Dr. Keith Cruise for his erudite observations regarding statistics and methodology. Thanks to Jennipher Roman for her generous help with data entry. Many thanks to friends and family members who provided support over the duration of this project. Finally, I would like to thank my wife, Cathy H. Bender, who tolerated my many nights at the computer with grace and kindness and assisted me in myriad ways throughout the process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER I

INTRODUCTION

Patient cooperation and optimal effort are essential to the validity of neuropsychological test results because most neuropsychological tests are normed on honest participants trying their best. Motivation to perform poorly is likely to produce grossly inaccurate results. Thus, the accurate detection of nonoptimal, exaggerated, or malingered performance is fundamental to the utility of these measures (McCaffery, Williams, Fisher, & Laing, 1997). The problem of malingered performance in neuropsychological assessment has received considerable attention lately regarding the magnitude of the problem and its detection.

Researchers have proposed numerous strategies and techniques to detect malingering.[1] These methods include the examination of atypical performance both within and between neuropsychological tests, documentation of unusual or incongruent behavioral presentation, and the use of specific tests of feigning. Most tests designed specifically to detect feigning have compared performance curves, frequency of failure on easy items, the magnitude of errors, or improbable response times (see Nies & Sweet, 1994). Despite a steady increase in the number of studies, data do not consistently

[1] For simplicity, the term "neurocognitive feigning" will be used to describe the production of poor performance on measures of neurocognitive function (a) in the absence of neurocognitive impairment or (b) beyond that which would be expected following a head injury. "Malingering" will be used to describe similar behavior when an external incentive is clearly identifiable.

support the use of any one measure over another. Rather, reliance on multiple measures and strategies is strongly recommended (Sweet, 1999). This introduction investigates the usefulness of individual detection strategies and proposes an integration of strategies to better discriminate feigned from non-feigned performance. The introduction consists of four sections that track and critique the development of neurocognitive feigning assessment.

The first section of the introduction explores the development of forensic neuropsychological assessment. The observed increase in research on malingering is at least partially due to an increased awareness of the potential for feigning in forensic contexts on both psychological and neuropsychological tests (Martin, Bolter, Todd, Gouvier, & Niccolls, 1993; Wiggins & Brandt, 1988).

The second section examines differential diagnosis and classification of disorders associated with malingering. Specific criteria that differentiate malingering from related disorders have limitations on conceptual and diagnostic grounds. An example of a diagnostic problem lies in the differentiation of Factitious Disorder from malingering. Another potential problem lies in the concept of secondary gain, which is said to be a driving force behind the dissimulation in malingering, but not in Factitious Disorder (for a review, see Rogers & Reinhardt, 1998). However, establishing what constitutes secondary gain and whether it exists is often difficult. In light of these difficulties, a brief summary of some of the conceptual problems associated with current differential diagnoses is provided.

In the third section, current research methods for the detection of neurocognitive feigning in neuropsychological evaluations are examined and critiqued. Most research employs one of three designs (simulation, known-groups comparison, and differential prevalence) and is usually limited to the study of a single detection strategy. A summary of current research implementing these designs and strategies is provided.

The introduction concludes with a fourth section that proposes a new measure of neurocognitive feigning and describes the research questions analyzed in the study. Each question evaluates important indicators of the validity of the currently proposed test: The Test of Cognitive Abilities (TOCA; Rogers, 1996). Specifically, the research questions examine the utility of the strategies employed by the TOCA, used individually and in combination, for the detection of neurocognitive feigning.

Forensic Clinical Neuropsychology

Neuropsychology has developed from the long-standing convention in neuroscience that the structure and function of neurons determine behavior (Kolb & Wishaw, 1996). Clinical neuropsychology is defined generally as the study and application of data regarding the relationship between brain function and behavior. The dual emphasis on behavior and neuroscience has put neuropsychology in a unique position to provide information about how brain damage affects behavior. Specifically, because neuropsychology focuses both on neural substrates of behavior and psychometrics, neuropsychologists can offer in-depth, quantified, normative data on brain-behavior relationships. Historically, available data were not objective and were

based only on opinions derived from interview and mental status examinations (Martell, 1992).

Neuropsychologists are increasingly asked not only to evaluate neurocognitive status but also comment on the validity of the results, with particular respect to the possibility of malingering. This is especially true in forensic settings. Both civil and criminal courts are increasingly interested in the behavioral consequences of brain injury and numerous studies have documented the increasing role of neuropsychologists in brain injury litigation (see McCaffery et al., 1997). Neuropsychological assessments are particularly important in cases of mild traumatic brain injury (mTBI) where sequelae (i.e., the results of pathology) are often not detected by neuroradiologic techniques. In such cases, neuropsychological evaluation may provide the primary source of evidence to support the claims of injury and need for financial compensation. The promise of large monetary rewards can increase the prevalence of injury exaggeration or the complete fabrication of symptoms (Binder & Rohling, 1996). Thus, the neuropsychologist is frequently asked to draw conclusions regarding the client's motivation to perform his or her best.

Most neuropsychological assessments are considered to be adequately valid. However, without the patient's optimum effort, the results may incorrectly indicate the presence of a deficit. For many years poor effort was not considered a problem because neuropsychological assessments were thought to be impervious to faking (Heaton, Smith, Lehman, & Vogt, 1978). However, it has become clear in recent years that neuropsychological assessments, like other types of assessments, are susceptible to the

effects of poor motivation, altered response styles, and outright faking. As a result, the extent to which neuropsychologists can identify feigned performance has been questioned and critiqued (see Nies & Sweet, 1994; Rogers, Harrell, & Liff, 1993; Sweet, 1999 for reviews), especially in the forensic arena (Arnett, Hammeke, & Schwartz, 1995; Bernard, 1990; Guilmette, Hart, & Giuliano, 1993).

The potential for dissimulation was demonstrated in Binder and Rohling's (1996) meta-analysis on the impact of financial incentives on disability, symptoms, and objective findings following closed head injury. After examining 18 studies with over 2,350 participants, the authors concluded that the most "impaired" test results were found in those patients with an incentive to feign (Cohen's $\underline{d} = 0.47$). They warn that clinicians must consider the individual client's financial incentive to perform below optimal levels when evaluating neurocognitive functioning. However, clinicians must be careful not to assume the client is malingering simply because there is a potential financial incentive to perform poorly.

Van Gorp et al. (1999) recently examined neurocognitive profiles of two groups of litigants with financial incentives. Prior to undergoing a comprehensive neuropsychological battery, the investigators divided the groups into those who "failed" symptom validity tests and those who did not. The authors concluded that relying on indices of feigning derived from standard neuropsychological tests did not accurately identify those litigants who failed the symptom validity tests. Specifically, the authors used a discriminant function analysis to study patterns of performance on the measures and found that the tests did not effectively differentiate the group of "suspected

malingerers" from other claimants. The authors argued that tests designed specifically to detect feigning are necessary in neuropsychological assessments.

Neuropsychologists often neglect feigning in their extensive test batteries. For instance, Lees-Haley, Smith, Williams, and Dunn (1996) in their study of 100 experts in neuropsychology, found that all had failed to include an objective measure of neurocognitive feigning in their forensic neuropsychological assessments. A possible explanation for this finding is that there is an assumption that malingering is not a fundamental concern, unless it is stipulated in the survey (or by the referral question). Also, some neuropsychologists use indices derived from existing neurocognitive measures and may not have listed them separately as malingering measures. The authors concluded that neuropsychologists underestimate the importance of malingering detection.

The fundamental importance of malingering detection has been underscored by Binder and Willis (1991), who have argued that the assessment of malingering is as important as any neurocognitive domain and should be assessed in every case where there is potential for financial gain. Similarly, Rogers (1997) has reminded clinicians to formally assess malingering in all situations in which negative consequences may be avoided.

For accurate identification of malingering to occur, alternative explanations for atypical responses must be ruled out. For example, anxiety and depression may result in involuntary fluctuations in performance on a variety of neurocognitive tests. Fatigue often reduces processing speed, which is a highly sensitive indicator of neurocognitive

impairment. Finally, nonconscious motivation to exaggerate symptoms and peculiarities of an individual's own concept of illness can produce unusual or lower than expected performance (Frederick, Sarfaty, Johnston, & Powel, 1994). Without differentiating malingering from these other explanations, any conclusions about feigning are equivocal.

Classification and Differential Diagnosis

Malingering

In this section, current classifications of malingering and disorders related to feigned presentations are discussed. Specifically, how the features of Factitious Disorder, Conversion Disorder, and Somatization Disorder can complicate differential diagnosis is examined.

The American Psychiatric Association (APA, 2000) defines malingering in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition - Text Revision (DSM-IV-TR) as, "intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military conscription or duty, avoiding work, obtaining financial compensation, evading criminal prosecution, obtaining drugs, or securing better living conditions" (p. 683). Thus, motivation to malinger can come from the desire for undeserved reward (e.g., money) or avoidance of adverse conditions (e.g., conscription).

The DSM-IV-TR (APA, 2000) classifies malingering as a V-code or, "other conditions that may be a focus of clinical attention," (p. 675) and offers a composite index of suspicion for malingering. Specifically, clinicians are encouraged to strongly suspect malingering when two or more of the following are present: (a) medicolegal

context of presentation, (b) marked discrepancy between claimed stress or disability and the objective findings, (c) lack of cooperation during evaluation and treatment, and (d) diagnosis of Antisocial Personality Disorder. Rogers (1990) reported that the presence of two or more of the above indicators correctly classifies about two-thirds of malingerers. However, for every malingerer correctly identified, nearly four genuine patients were misclassified as malingerers. This information suggests that using the DSM-IV-TR criterion of two or more indicators is lacking in positive predictive power (i.e., there is only a 20.1% likelihood of being correctly labeled a malingerer), and is only marginally sensitive (i.e., 67% of malingerers are identified as such).

The consequences of being falsely identified as a malingerer can be devastating. Thus, the number of false positives (i.e., the number of genuine patients incorrectly classified as malingerers) a measure yields is especially important to researchers, clinicians, and patients. With regard to the DSM-IV-TR suspicion index, the false positive rate is very high. Relying solely on this index, or on any measure with a similar false positive rate, may be considered a deficient standard of practice. A more detailed discussion of utility estimates in neurocognitive feigning detection is found in later sections of this chapter.

Factitious Disorder

Factitious Disorder, like malingering, lacks a clear organic etiology for its symptoms, and is presumably under conscious control of the individual (APA, 2000). It differs from malingering in that symptoms, though produced intentionally, are due to a psychological need to assume a sick role. In contrast, the motivation behind malingering

is to obtain an identifiable external gain (e.g., money). The DSM-IV-TR (APA, 2000) divides Factitious Disorders into Factitious Disorder with Predominantly Psychological Signs and Symptoms (FDPS) and Factitious Disorder with Predominantly Physical Signs and Symptoms. The latter diagnosis subsumes "Munchausen's syndrome," which typically involves a long history of hospitalizations and medical contacts. When negative results are obtained, new symptoms often appear in these patients. When confronted with clearly incongruent findings or contradictions, the patient may threaten litigation and usually seeks a second or third opinion (Cullum, Heaton, & Grant, 1991).

The essential feature of FDPS is the presence of intentionally produced psychological symptoms without external incentive. However, it may be quite difficult to assess the degree of voluntariness of the behavior. Jonas and Pope (1985) argued that the extent to which behavior is consciously or voluntarily motivated is on a continuum rather than a categorical variable. They provided evidence that factitious disorders, somatoform disorders, and malingering are remarkably similar in age of onset, course, responsiveness to treatment, and comorbidity of personality disorders. The findings of Rogers, Bagby, and Vincent (1994) support the above findings. They found that FDPS patients and malingerers were very similar in symptom presentation, thus further questioning the diagnostic legitimacy of both disorders.

The presence of an external incentive is said to differentiate malingering from Factitious Disorder (APA, 2000). However, as Cunnien (1997) noted, "this criterion runs contrary to knowledge about the multilevel nature of behavior. The mere presence of external gains cannot negate the primacy of psychological motives" (p. 25). For example,

a “malingerer” may feign illness to alleviate anxiety or to avoid conflict with authorities rather than feign to achieve external compensation. As such, this behavior is not accurately encapsulated by either DSM-IV-TR (APA, 2000) classification. Much of the relative importance of internal and external motivation is subjectively decided by the examiner. Rogers, Bagby, and Rector (1989) provided evidence that the diagnostic legitimacy of FDPS was seriously limited. They contended that FDPS should not be included in the next DSM (APA, 1994) due to the poor inclusion, exclusion, and outcome criteria for the disorder.

An additional problem with the DSM-IV-TR (APA, 2000) classification of malingering is the specification that the presence of other conditions involving faked or exaggerated symptoms rules out malingering. In other words, comorbidity of malingering and Factitious Disorder (for example) is impossible. This distinction fails to recognize the fact that psychological and financial incentives can often coexist and behavior can be motivated by both external and internal gain (Slick, Sherman, & Iverson, 1999).

Secondary Gain

The term secondary gain is typically used pejoratively to describe essentially manipulative acts aimed at securing attention, special services, or external incentive. However, Rogers and Reinhardt (1998) noted that there are at least three different models of secondary gain. The psychodynamic model conceptualizes secondary gain as involuntary and the behavioral model views it as caused by others. If accurate, these models undermine the criteria that the pursuit of secondary gain must be volitional. Each model emphasizes a different motivation for gain and as a result contributes to a lack of

consensus regarding its conceptualization. In light of the conceptual heterogeneity, the authors cautioned that use of the term secondary gain is likely to be misunderstood and therefore should be avoided. Overall, they concluded that secondary gain represents an ambiguous constellation of constructs that when used inappropriately can threaten a patient's integrity. Attempts to classify disorders in which incentive and motivation are essential features are impeded by this poorly understood yet commonly referenced construct.

Ganser Syndrome

Ganser syndrome is a rare condition in which the patient's symptoms can simulate mixed organic, affective, and psychotic illnesses. First diagnosed by Ganser in 1898, it was described as involving clouded consciousness, hallucinations, conversion symptoms, amnesia, and "approximate answers." Contemporary theorists now suggest that Ganser Syndrome is typified by the presence of approximate answers alone, or a portrayal of neurocognitive impairment in which individuals nearly miss the correct response (e.g., claiming there are 65 minutes in an hour; Heron, Kritchevsky, & Delis, 1991). This pattern is exhibited despite an intact memory for less overlearned material.

Formerly considered a subtype of FDPS, Ganser Syndrome was reclassified as an associated feature of dissociative amnesia and dissociative fugue in DSM-IV (APA, 1994). It differs from malingering in that the Ganser Syndrome patient cannot suppress his or her performance and does not appear to have clear goals for the performance (Heron et al., 1991). It is not clear whether or not symptom production is intentional. An additional complication to accurate diagnosis is the finding that approximate answers are

not limited to Ganser Syndrome, and have been documented in traumatic, infectious, toxic, and neoplastic insults. To date, very few experimental studies have involved neuropsychological correlates of Ganser Syndrome and (based on this literature review) apparently none have attempted to differentiate Ganser Syndrome from malingering and other clinical conditions.

Somatoform Disorders

The two most relevant Somatoform Disorders to consider in differential diagnosis involving neurocognitive feigning are Somatization Disorder and Conversion Disorder (Cullum et al., 1991). Both disorders are defined by: (a) the presence of symptoms that suggest a medical condition but are not fully explained by the condition, and (b) the absence of voluntary control over symptomatology (APA, 2000). Both disorders resemble malingering in that symptoms are usually a result of processes other than those proposed by the patient. Somatization Disorder, also known as "Briquet's syndrome," is most frequently diagnosed in females and involves frequent medical interventions. The complaints associated with Somatization Disorder typically have a dramatic, attention-seeking quality. These individuals often consult with general medical practitioners rather than psychologists for at least two reasons: (a) they often have genuine disorders coexisting with the Somatization Disorder; and (b) they believe their somatic symptoms are medically, rather than psychologically, based. Prevalence rates of Somatization Disorder are difficult to estimate but most suggest that it is relatively rare. However, Maxmen and Ward (1995) suggested that it could be as high as 14% among psychiatric patients referred from medical/surgical units. Somatization Disorder also is much less

frequently diagnosed by non-physicians (APA, 2000). The latter point may be important to psychologists because psychologists rarely see persons with Somatization Disorder during their careers; differential diagnoses is likely to be more difficult in such rare cases.

The essential feature of Conversion Disorder is the presence of symptoms or deficits affecting neurological function that are judged to be associated with psychological factors (APA, 2000). The symptoms are not intentionally produced or feigned. Conversion Disorder and Somatization Disorder differ in at least two respects. Conversion Disorder typically involves one symptom, whereas Somatization Disorder often presents with many symptoms in numerous systems. Conversion Disorder also tends to arise quickly while Somatization Disorder emerges gradually (Maxmen & Ward, 1995). However, the disorders are similar in that bona-fide neurological problems (e.g., head injury), which frequently involve nonspecific symptomatology, often accompany both disorders (McCaffrey et al., 1997).

Conversion Disorder in particular may be quite difficult to differentiate from medical illness. Common misdiagnoses include myasthenia gravis, idiopathic dystonia, and multiple sclerosis. The presence of a neurological illness does not preclude diagnosis of Conversion Disorder and about one-third of individuals with conversion symptoms have a current or prior neurological condition (APA, 2000). Malingerers, especially those who are well-coached, may present with the same symptoms. The sole distinguishing feature between Conversion Disorder and malingering may be the absence of intentional symptom production. As mentioned, the question of the degree to which symptom production is conscious may be extremely difficult to answer.

In summary, accurate classification of malingering can involve complex differential diagnoses. The overlap of behaviors associated with malingering and other disorders complicates diagnosis. The heterogeneity of other presentations resembling malingering militate against confident classification.

Other Factors Related to Diagnostic Confidence

Certitude of Malingering Classification

Using current methods, the classification of malingering is not absolute. It is subject to the clinician's judgments and his or her confidence in the findings. Rogers (1997) argued that when evidence of malingering is not unequivocal, it should be considered as "unreliable data," rather than malingering. When data are reliable, clinicians are encouraged to use five levels of certainty when making conclusions based on tests of feigning. The highest level is termed, "definite" meaning that the test accurately classifies 90% or more of feigning and nonfeigning individuals. The next level of certainty is "probable," meaning that research has shown it to classify individuals accurately at least 75% of the time. A "tentative" level of certainty occurs when research is statistically significant (i.e., produces group differences) but lacks practical value (i.e., poor classification rates). The fourth and fifth levels of certainty lack empirical validation and are labeled "speculative" and "unsupported," respectively. In subsequent sections of this proposal, these levels are referenced when discussing the findings of recent neurocognitive feigning research.

Problems of Prevalence

An important issue that affects the level of classificatory confidence is that of prevalence (i.e., rates of occurrence). The prevalence rate (also commonly known as base rate) of a given condition is defined by the number of cases with the condition divided by the number of individuals in the population (Gouvier, 1999). When the prevalence is very low, other explanations for the presentation are often more likely.

Base rates have been shown to influence classification accuracy in predictable ways. In differential diagnosis, they affect diagnostic accuracy by reducing the accuracy of prediction of the rarer condition and potentially boosting the accuracy of prediction of the more common condition (Gouvier, 1999). Thus, base rates are important in test development because tests, though reliable and valid, may not contribute to diagnostic accuracy beyond that expected with the base rate alone. The relative advantages and disadvantages of utility estimates that take into account base rates will be addressed in the next section.

According to the DSM-IV-TR (APA, 2000), the prevalence rate of Factitious Disorder is not known; it states simply that Factitious Disorder appears to occur more often in males than in females. Sutherland and Rodin (1990) reported that of 1,288 medical inpatients referrals for psychiatric consults, less than 1% met criteria for any Factitious Disorder and only one was diagnosed with FDPS. Overall, diagnosis of Factitious Disorder remains rare and controversial (Rogers et al., 1989). Likewise, the estimated prevalence rates in the general population are quite low for Somatization Disorder (0.2 - 2%) and Conversion Disorder (1 - 3%; APA, 2000). Though somewhat

more prevalent in psychiatric and neurologic clinics, these disorders are still considered rare. Thus, they may be less likely to be diagnosed with the same level of confidence as more common disorders.

Accurate prediction of malingering is likewise hindered by the lack of information regarding its prevalence. In a survey of 320 forensic psychologists, Rogers, Sewell, and Goldstein (1994) estimated that malingering occurred in 15.7% of forensic and 7.4% of nonforensic cases. Rogers, Sewell, Salekin, and Goldstein (1998) found similar percentages (17.4% and 7.8%, respectively) in a subsequent survey.

Prevalence estimates of neuropsychological malingering are even less well-documented. Rogers et al. (1993) in a review of the available literature, offered very preliminary data to suggest that as many as 50% of personal injury claimants may be feigning neurocognitive impairment. Other estimates range from 20% (Griffin, Normington, May, & Glassmire, 1996) to 67% (Heaton et al., 1978) depending on the setting. Greiffenstein, Baker, and Gola (1994) found that 41% of 106 consecutive referrals for neuropsychological evaluations of mild traumatic brain injury (mTBI) met 2 of their 4 criteria for malingering. Sweet (1999) reported that a comprehensive review of the available estimates revealed an average prevalence rate of approximately 15% in all neuropsychological evaluations across settings. A possible explanation for the variability in estimates is that some studies relied on anomalous tests results to indicate malingering without establishing the presence of poor motivation and clear external incentive.

Utility Estimates for Feigned Neurocognitive Impairment

Diagnostic confidence depends on how well a given test correctly classifies groups (e.g., malingerers vs. genuine patients) in light of the condition's prevalence rate. The most commonly used measures of diagnostic utility (i.e., utility estimates) are defined below with reference to neurocognitive feigning (see also Table 1):

1. True Positives: The proportion of individuals correctly identified as feigning neurocognitive deficits.
2. True Negatives: The proportion of individuals correctly identified as not feigning.
3. False Positives: The proportion of individuals incorrectly identified as feigning
4. False Negatives: The proportion of individuals incorrectly identified as not feigning.
5. Sensitivity Rate: The proportion of individuals who are feigning identified as such by the test. It is calculated as the number of true positives divided by the total number of individuals actually feigning.
6. Specificity Rate: The proportion of individuals who are not feigning identified as such by the test. It is calculated as the number of true negatives divided by the total number of individuals not feigning.
7. Positive Predictive Power (PPP): The probability that an individual identified by the instrument as feigning actually is feigning. PPP is calculated as the number of true positives divided by the total number of positive predictions.

8. Negative Predictive Power (NPP): The probability that an individual identified by the instrument as not feigning actually is not feigning. NPP is calculated as the number of true negatives divided by the total number of negative predictions.
9. Hit rate: The total number of correct classifications divided by the total number of classifications. It is calculated as the number of true positives plus true negatives divided by the total number of predictions.

Table 1

Utility Estimates Commonly Used in Diagnostic Validity Applied to Feigning

	Gold Standard				
	Feigning		Not Feigning		
Measure					
Feigning	True Positives (a)	+	False Positives (b)	=	Total positive predictions
Not Feigning	False Negatives (c)	+	True Negatives (d)	=	Total negative predictions

Note. Positive Predictive Power = a / a + b; Negative Predictive Power = d / c + d; Sensitivity = a / a + c; Specificity = c / a + c; Hit Rate = a + c / a + b + c + d.

Sensitivity and specificity are the most commonly reported estimates in studies involving the development of diagnostic instruments. Researchers have noted an inverse relationship between the two estimates: maximizing one frequently limits the other (Anastasi, 1997; Kazdin, 1992). When using diagnostic instruments, clinicians face a

dilemma of whether to rely more heavily on sensitivity or specificity. Situations in which the main concern is not to miss a diagnosis at the cost of over-inclusiveness (e.g., screening), the sensitivity of the measure is emphasized. Conversely, in situations requiring more certainty about ruling out another possible explanation, specificity is stressed. Neither sensitivity or specificity take into account the base-rate of malingering.

Binder (1995) has argued that PPP and NPP are better estimates of clinical utility because they take into account base rates. For example, PPP provides a more clinically useful estimate because it reflects the probability that a patient is feigning given a positive result. In contrast, sensitivity reflects the probability of a positive result given that an individual is feigning; the clinician's task actually is to make the opposite determination. A counter-argument is that PPP and NPP are less stable estimates due to their dependence on base rates and are consequently less useful. Some researchers have begun to publish both types of estimates. In order to understand the implications of utility estimates, it is important to understand the strengths and weaknesses of the research designs used in malingering research.

Research Designs for the Study of Feigning

No gold standard or "true measure" has been established to which to compare an instrument's validity as a means of detecting neurocognitive feigning (Cercy, Schretlen, & Brandt, 1997; Rogers et al., 1993). Therefore, researchers typically turn to one of three research designs to investigate neurocognitive feigning: known-groups comparison, simulation, and differential prevalence.

Known-Groups Comparisons

Known-groups comparisons consist of two stages: (a) identifying criterion groups and (b) analyzing similarities and differences between the groups (Kazdin, 1992). The primary advantage of known-groups comparison is its clinical relevance. For instance, known malingerers are typically found in clinical settings (e.g., hospitals) and faced with difficult circumstances. Their motivation to malinger is based on real-world issues with clear consequences for unsuccessful feigning (Rogers, 1997). A fundamental difficulty with known-groups design is in the accurate classification of the criterion groups (i.e., establishing accurate groups of known malingerers and known brain-injury patients). This task typically requires expert opinion and thorough evaluations; such extensive efforts can discourage the use of the design. A related limitation of known-groups comparison is that malingerers are rarely identified as such (Cercy et al., 1997), and are limited in availability for research as a result.

Simulation Design

Simulation (i.e., analogue) studies involve a quasi-experimental design with three groups: (a) simulators (i.e., normal individuals portraying a feigned injury), (b) controls (i.e., normal individuals under standard instructions), and (c) a clinical comparison sample (i.e., genuine patients performing their best). Group performance is then compared; if reliable differences in performance are found, then the test is likely to discriminate feigners from controls and genuine patients. Simulation designs suffer from limited external validity. The extent to which simulators reflect actual malingerers is unknown. The strength of the design lies in its experimental rigor and internal validity.

The consequences for unsuccessful malingering in the "real-world" can be quite different from those encountered in an experiment. In actual court cases, failing to fake neurocognitive impairment convincingly may not only result in nonpayment but also may result in fines or imprisonment if perjury is committed. However, most studies of malingering do not include negative incentives. Rogers and Cruise (1998) investigated the notion that real-world incentives often involve negative consequences. Simulators given negative incentive (i.e., public posting of their poor performance) were more believable in their efforts on the Structured Interview of Malingered Symptoms (SIMS; Smith, 1992) than those given positive incentive (course credit and the opportunity to win $50). The authors concluded that simulation designs should attempt to include negative incentives analogous to those found in real-world settings.

Some clients stand to gain substantial rewards if they appear impaired in certain settings (e.g., disability determinations). McCaffery et al. (1997) have argued that because of these rewards, the prevalence of coaching clients to appear impaired is increasing in forensic settings.[2] In light of this observation, simulation research has recently employed coaching conditions to increase the applicability to coached feigners (see Nies & Sweet, 1994). Rose, Hall, Szalda-Petree, and Bach (1998) found that coached simulators were more difficult to detect than were noncoached simulators. Franzen and Martin (1996) and Martin et al. (1993) confirmed this finding on similar neuropsychological tests.

Differential Prevalence Design

The third design employed in neurocognitive feigning research is differential prevalence. The design works on the assumption that the prevalence rates of malingering are different across settings and referral questions. This assumption stems from the idea that persons in certain settings have potentially large incentives that will increase their likelihood to malinger (Rogers, 1997). For example, due to the possibility of large monetary rewards, individuals in litigation often have a potential incentive to exaggerate or modify symptomatology. Thus, they have been used as the "more-likely-to-malinger" group to which clinical groups with less potential incentive can be compared (e.g., Binder & Willis, 1991; Millis, 1992). For instance, Millis (1992) found that a small number of mild head-injured participants who were seeking financial compensation scored significantly lower on the Recognition Memory Test (RMT; Warrington, 1984) than did moderate to severe head injury participants not seeking compensation. The flaw in differential prevalence research is that persons seeking compensation may be more impaired in either neurocognitive or psychological functioning than those who do not seek litigation.

Several studies have attempted to define a subset of potential malingerers by dividing litigants by performance on measures of malingering. A relatively consistent finding is that some patients involved in litigation perform poorly both on tests of neurocognitive feigning and on memory assessments, when compared to patients not in

[2] Coaching typically involves providing details to the participant about how to fake convincingly without being detected.

litigation (Binder & Willis, 1991; Millis, 1992; Trueblood, 1994; Youngjohn, Burrows, & Erdal, 1995). However, it may be inappropriate to conclude that malingering is responsible for this finding. First the prevalence rate of malingering in this sample is simply assumed and not empirically measured. Second, a diagnosis of mild head injury may be neurologically, but not psychologically, accurate. Motivation to perform may be compromised if the individual perceives the injury to be worse than the diagnosis suggests. For example, an artist who suffers a minor head injury but experiences reduced fine motor ability from it may perceive the injury to be more devastating than someone less dependent on fine motor ability. Thus, the artist may be more willing to accentuate problems if she believes her complaints will be minimized otherwise (by way of only mild impairment ratings on the neuropsychological battery).

Weissman (1990) addressed the issue of perceived level of impairment with respect to the forensic setting. He argued that, if symptoms improve following resolution of a court case, it should not necessarily be considered evidence of malingering. Alternative explanations for poor performance during the assessment must be considered. As a consequence of protracted litigation, morale may be poor, treatment effectiveness may be reduced, and the probability of iatrogenic conditions may increase. Thus, the patients' perception of their problems may improve following the conclusion of the case.

Although differential prevalence design appears logical, the overall interpretability of its results is problematic and largely unprofitable. Rogers (1997) cautions that it lacks practical value because differences in prevalence rates are inferred, not measured. No systematic data from multiple sources indicate that malingering is

probable (e.g., occurring in more than 75% of the sample) in any setting, including litigation. So, constructing a group based on the notion that malingering is characteristic of persons in litigation is misleading and apt to be misunderstood.

Strategies for Detection of Neurocognitive Feigning

Investigations of specific feigning detection strategies is a relatively new pursuit. In contrast, a great deal of research has yielded a variety of indices and techniques that have little or no reference to theory-driven strategies. More recent studies emphasizing strategic detection have produced highly effective means of assessing feigned neurocognitive performance.

Rogers et al. (1993) proposed six strategies for detecting feigned performance on neuropsychological assessments based on previous research. They described the following strategies: (a) Symptom Validity Testing (SVT), (b) Floor Effect, (c) Magnitude of Error, (d) Performance Curve, (e) Atypical Presentation, and (f) Psychological Sequelae. Table 2 displays these strategies and gives brief descriptions of each. Table 3 lists research designs commonly used to validate tests employing these strategies. Although not expressly addressed by Rogers et al. (1993), the utility of forced-choice testing (FCT) for detecting neurocognitive feigning also has been investigated and will be included in the following review.

Symptom Validity Testing

SVT utilizes a multiple-choice procedure in which performance is compared to chance probabilities. It differs from FCT (see subsequent section) in that comparisons are made to chance performance rather than to the performance of other groups. The

Table 2

Detection Strategies for Neurocognitive Feigning

Strategy	Description
Floor Effect	Failure on very simple items that the vast majority of impaired persons get correct suggests feigned performance
Performance Curve	Performance consistent with difficulty of item; curves with a positive slope as item difficulty increases suggest feigning
Symptom Validity Testing (SVT)	Comparison of multiple-choice performance to chance probability; performance significantly below chance level is considered unequivocal evidence of feigning
Forced-Choice Testing (FCT)[a]	Comparison of multiple-choice performance to other group performance; performance significantly below clinical groups is considered suggestive of feigning
Magnitude of Error	The extent to which a response is incorrect; patterns of gross and near misses. Endorsement of highly unlikely errors may indicate feigning.
Atypical Presentation	Performance inconsistent with norms or expected patterns can suggest feigning; can be measured within or between tests
Psychological Sequelae	Subjective complaints associated with impairment that are inconsistent with diagnosis or are exaggerated can indicate feigning

Note. Adapted from Rogers et al. (1993)
[a] Described in Rogers and Vitacco (in press)

Table 3

Detection Strategies, Research Designs, and Representative Measures of Feigned Neurocognitive Impairment

Strategy	Research Designs		Representative Measures
FCT	Sim	D-P	PDRT, HDMT
SVT	Sim	D-P	PDRT, HDMT
Performance curve	Sim	D-P	VIP, DCT, DCT
Floor effect	Sim	D-P	FIT, TOMM, CARB
Magnitude of error	Sim	D-P	WMS-R
Response time	Sim	D-P	MAS, PDRT, DCT
Atypical Presentation		D-P	WMS-R, HRB

Note. Adapted from Rogers (1997). FCT = Forced-Choice Testing; SVT = Symptom Validity Testing; Sim = Simulation design; D-P = Differential prevalence design; PDRT = Portland Digit Recognition Test; HDMT = Hiscock Digit Memory Test; VIP = Validity Indicator Profile; DCT = Dot Counting Test; FIT = Fifteen Item Test; TOMM = Test of Malingered Memory; WMS-R = Wechsler Memory Scales-Revised; MAS = Memory Assessment Scales; HRB = Halstead-Reitan Neuropsychological Test Battery.

rationale is that significantly below-chance performance suggests that individuals taking the test were aware of the correct response but chose to respond incorrectly. Brady and Lind (1961) introduced two-choice multiple-choice testing to examine a case of hysterical blindness. With each stimulus alternative presented 50% of the time, the probability of correct responding by purely guessing is 50%. They found the individual's performance was significantly below chance and concluded that the patient was faking.

One major advantage of SVT over other strategies is the lack of other plausible explanations for below-chance performance. Its main limitation is the small percentage of simulators who score below chance (Rogers et al., 1993). Martin et al. (1993) found that only 46% of naïve simulators and 10% of sophisticated (i.e., coached) simulators achieved scores below chance levels, suggesting that SVT is not sufficiently sensitive to warrant its use in some professional settings. As will be discussed, SVT appears to suffer from the converse problem of FCT (i.e., SVT has excellent specificity but lacks sensitivity, while FCT has good sensitivity but poor specificity).

Much of the support for SVT has accrued through case studies (Binder, 1992b; Brady & Lind, 1961; Pankratz, 1979). Its clinical usefulness in group settings has been less well documented. For example, The Hiscock Digit Memory Test (HDMT; Hiscock & Hiscock, 1989) was developed as a modification of previous multiple-choice procedures and has been administered in experimental designs. In short, a 5-digit number is printed on a card or presented on a computer screen (see Martin et al., 1993). Following a brief delay, examinees are presented with two 5-digit numbers and are asked to identify the correct alternative. Delays are set at 5, 10, and 15 seconds to increase the appearance of increased gradations of difficulty. Slick, Hopp, Strauss, Hunter, and Pinch (1994) found that only 15% of a simulating group scored significantly below chance on the HDMT while all members of a traumatic brain-injury group scored at or above chance. Based on these results, the authors suggested that the HMDT was not a useful discriminator when performance was compared to chance. In contrast, they espoused the

HDMT as a FCT where group performance can be compared (see also Hiscock, Branham, & Hiscock, 1994).

Researchers have investigated the usefulness of the forced-choice 72-item Portland Digit Recognition Test (PDRT; Binder, 1993; Binder & Willis, 1991) as an SVT of feigned neurocognitive deficits. The PDRT is very similar to the HDMT (e.g., 5-digit numbers are presented and the participant chooses from 2 choices presented shortly thereafter); an important difference is the addition of a distraction task and longer delays. In a differential prevalence design, Binder (1993) found that none of the moderate to severe head-injured patients not seeking litigation scored below chance; whereas 17% of a group of mild head-injured patients in litigation scored below chance. He concluded that (a) when performance is below chance probability, SVT is a convincing feigning detection strategy, and (b) the presence of financial incentive explained the differences in performance. However, the reliance on the differential prevalence design confounds conclusions about the likelihood of feigning by assuming that the level of genuine impairment is not increased in litigating groups.

Excellent specificity and inadequate sensitivity have been documented on other multiple-choice tests of feigning. Tombaugh (1997) and Frederick and Foster (1991), for example, have noted very few suspected malingerers perform below chance on the TOMM and the TONI, respectively. As previously noted in the current review, coaching likely reduces the number further. In terms of Rogers (1997) classifications of diagnostic certitude, the results from studies using the SVT strategy range from Unsupported (e.g.,

when no cases of actual or simulated feigning are identified) to Definite (e.g., when an individual is identified as feigning, it is unequivocal).

Forced Choice Testing (FCT)

Although the term FCT is frequently used interchangeably with SVT, an important difference between the two procedures exists. FCT compares group performance (e.g., simulators vs. brain-injured), whereas SVT compares performance to chance probabilities. They are similar in that both measure a specific ability (typically recognition memory) by presenting a large number of items in a multiple-choice format. The rationale for FCT lies in the assumption that decision rules can provide cutting scores to differentiate groups. For example, based on comparisons for 110 patients with unambiguous brain dysfunction, Binder (1993) reported that scores below 39 correct (i.e., 54.2%) on the PDRT effectively differentiated the groups and were considered indicative of incomplete effort.[3]

Discontinuation rules for the PDRT have been employed when an individual performs well. This is important because the PDRT takes approximately 45 minutes to administer and because many of the patients who take the PDRT do not show evidence of poor motivation (Binder, 1993). When the PDRT suggests questionable motivation, the test can be continued. Binder (1993) published this abbreviated form of the PDRT and

[3] Much of the available literature uses terms such as "nonoptimal effort," "poor motivation," "feigning," and "malingering" interchangeably. However, for the purposes of this review, unless it is clear that the design of a study in question explicitly measures malingering (e.g., a scenario is used and includes external incentive), it is assumed that the study is a measure of less specific constructs, such as feigning, motivation, or effort.

recommends using a conservative cutoff of 7 correct of 9 difficult items (i.e., 77.8%) as grounds for discontinuation (i.e., not feigning).

Guilmette et al. (1993) compared a brain injury group to a feigning group on the HDMT. They concluded that performance below 75% correct (i.e., 54 or fewer correct out of 72) strongly suggests exaggeration of symptoms rather than true brain injury. This cutting score correctly classified 100% of both the brain-injured and simulating groups. Sixty-four percent of the simulation group was misclassified as not feigning when SVT criteria were used.

The authors noted that the transparency of the HDMT might be a liability. Some simulators complained that the test was so easy they could not fake in a nonobvious manner.

Iverson, Franzen, and McCracken (1991; 1994) constructed the 21-Item Test as a measure of recall and as a FCT of feigning. The test consists of 21 words that are presented orally. Following a free recall trial, the patient is asked to identify the target word when it is presented with a foil (i.e., a 2-choice format). The studies to date have employed simulation designs and have yielded highly variable results. Depending on the cutting score and the composition of the clinical comparison group, classification rates vary from 20% to 80% (Frederick et al., 1994; Iverson & Franzen, 1996; Iverson et al., 1991, 1994).

According to Rogers' (1997) categorization, FCT falls within the Tentative range, meaning research has found significant differences but lacks adequate classification rates. Because FCT does not rule out as many possible explanations for the results as SVT (i.e.,

FCT is typically more sensitive than specific), the current estimates may over-estimate the utility of FCT in clinical settings (Etcoff & Kampfer, 1996; Nies & Sweet, 1994; Pankratz & Binder, 1997). In sum, research tends to support the use of the HDMT and PDRT in forensic settings; however, sole reliance on FCT fails to rule out the effects of psychiatric and neurological disorders on performance.

Floor Effect

The Floor Effect strategy assumes that failure on very easy items that most impaired patients answer correctly is unlikely to reflect a genuine deficit. For instance, Rogers et al. (1993, p. 260) reported that surveyed forensic experts viewed a client's failure on very simple informational items (e.g., "Which is bigger, a horse or a dog?") as indicative of malingering. The Rey Fifteen Item Test (FIT; Rey, 1964), the Computerized Assessment of Response Bias (CARB; Allen, Conder, Green, & Cox, 1997), and the Test of Memory Malingering (TOMM; Tombaugh, 1997) are examples of tests using the Floor Effect strategy that have received considerable attention in the neuropsychology literature. Only recently have researchers acknowledged these and similar tests as employing a specific conceptually-driven strategy.

Greiffenstein, Baker, and Gola (1994) evaluated the utility of a simple formula derived from the Digit Span subtest of the WMS-R (Wechsler, 1987). The "Reliable Digits" strategy was calculated by summing the raw scores for the longest string of digits repeated backward and longest string forward. As stipulated by the WMS-R, each item of the Digit Span subtest includes two trials; both had to be correct to qualify for inclusion in the formula. The authors determined that performance below 7 was very unusual for

patients and therefore represented a floor effect. They employed a differential prevalence design comparing severe TBI patients and probable malingerers. The investigators argued that their sample of litigating participants incorrectly assumed that auditory attention (i.e., Digit Span) was sensitive to brain injury and consequently scored worse than honest participants. In other words, they performed poorly on an easy task. The cutting score yielded adequate sensitivity (86.0%) but poor specificity (57.1%). Meyers and Volbrecht (1998) later cross-validated the study on a sample of mild TBI patients, approximately half of whom were in litigation. Consistent with the previous results, they found that litigating participants (whom the authors labeled suspected malingerers) consistently recalled significantly fewer digits than non-litigating participants. The cutting score of 7 or less correctly classified 95.9% of non-litigating clients. This score also classified 48.9% of litigating clients as feigning. The implications of the findings are limited by the possibility that litigants pursue litigation because they are more impaired. The tacit assumption that litigating participants are malingering rather than performing differently in response to any number of influences obscures interpretability. (see Rogers, 1997; Sweet, 1999; Weissman, 1990).

The Fifteen Item Test (FIT; Rey, 1964) is a frequently used screen of neurocognitive feigning that was introduced specifically as a means of validating memory complaints. The FIT deceptively appears to be difficult, but is very simple. Patients are given 10 seconds to memorize 15 items, grouped in three columns and five rows. This presentation facilitates encoding of 3 or 5 groups of stimuli rather than 15 separate bits of information. Typically, two recall scores are derived: (a) the total number

of correctly recalled items, and (b) the number of correctly ordered sets of items (i.e., rows or columns). Based on her clinical experience, Lezak (1995) suggested that only the most severely brain-damaged individuals fail to recall at least nine items on the FIT. Goldberg and Miller (1986) suggested that this cutting score may effectively discriminate mentally retarded patients' performance from feigners. However, Schretlen, Brandt, Krafft, and Van Gorp (1991) noted limitations to the Goldberg and Miller study: (a) the absence of a simulating group or a group of suspected malingerers, and (b) 37.5% of the of the mentally retarded participants performed below the cutting score. Other studies (Bernard, Houston, & Natoli, 1993; Guilmette et al., 1994) reported a lack of consensus regarding optimal cutting scores for neurocognitive feigning on the FIT.

Schretlen et al. (1991) demonstrated another problem with the proposed cutting scores for neurocognitive feigning on the FIT. They tested the FIT on a heterogeneous group of normal controls, severely mentally-ill patients, TBI patients, mixed neuropsychiatric patients, amnesics, suspected malingerers, and simulators of amnesia. The authors noted that the FIT lacked clinical utility because a single cutting score resulted in an extremely wide range of classification rates.

Research on the FIT also has found poor sensitivity but good specificity rates when classifying groups. Greiffenstein, Baker, and Gola (1994) obtained sensitivity of only 60.0% when discriminating between TBI and probable malingering groups. In contrast, specificity was 82.0%. Arnett, Hammeke, and Schwartz (1995) compared performance of neurological patients to simulators. Using a cutting score of two correct rows resulted in excellent specificity (96.1%), but very low sensitivity (47.4%).

Similarly, Millis and Kler (1995) found that using a cutting score of seven correct on the FIT lead to excellent specificity (100%), but poor sensitivity (56.6%) when discriminating between brain-injured and simulator groups. Given the good specificities reported in these studies, the FIT may serve as a valid method of screening out genuine cases of TBI.

Pachana, Boone, and Ganzell (1998) presented a case study that documents the potential limitations of using the FIT as a measure of feigning. A patient diagnosed with Wernicke-Korsakoff's syndrome was administered a full battery of neuropsychological tests, including measures of malingering such as the FIT. Consistent with the diagnosis, the patient yielded a profile of relatively intact intellectual functions and dense but circumscribed amnesia. According to the authors, there was no question of malingering, exaggeration, or poor motivation. The patient failed every item of the FIT, producing a "feigned performance" on the FIT. The authors cautioned that the FIT was "clearly contaminated by the presence of organic amnesia," and suggest that other tests of neurocognitive feigning be used instead (p. 22).

A relatively large amount of research has been conducted on the FIT. However, according to the levels of certitude offered by Rogers (1997), it falls into the Tentative category of classificatory accuracy. In other words, although group differences have been found, the classification rates tend to be lower than 75%. Thus, research does not support its use as a means of detecting feigned versus non-feigned performance (Etcoff & Kampfer, 1996). As a potential solution, Griffin, Glassmire, Henderson, and McCann (1997) investigated the utility of a qualitative scoring system with modifications to the

administration and presentation of FIT; they termed this new version the Rey II. The authors attempted to decrease the face validity, and thereby increase the discriminability of the FIT by replacing one row with one slightly more difficult row and putting boundaries around two rows in the stimulus. The Rey II yielded a sensitivity rate of 73.5% and specificity of 86.2%. The overall classification rate was 85.5%. Although their results provided incremental validity, other studies have not yet replicated their findings.

The Floor Effect strategy can also be used on multiple-choice tests. These tests allow for the development of quantitative indices of the likelihood of obtaining a correct response based on honest performance (e.g., ≥ 90%). Researchers have investigated the validity of memory tests using the Floor Effect on multiple-choice tests with promising results (Allen et al., 1997; Frederick & Foster, 1991; Hiscock & Hiscock, 1989; Tombaugh, 1997).

Research on the HDMT has also addressed its utility as a vehicle for the Floor Effect strategy. Guilmette et al. (1993) found that using a cutting score greater than or equal to 90% correct (i.e., 66 of 72 trials) on the Hiscock Digit Memory Test (HDMT; Hiscock et al., 1989) correctly classified 90% of a brain-damaged group and 90% of a simulating group. They concluded that performance below 90% (i.e., more than six errors) should raise suspicion in the clinician that "at the very least, motivation to do well may not be optimal" (p. 67).

Guilmette et al. (1994) also applied the Floor Effect strategy on a 36-item version of the HDMT. Using the same 90% cutting score (four or more errors), the test correctly classified 85.1% of simulators, 100% of brain damaged patients, and 94.5% of depressed

inpatients. Slick et al. (1994, 1996) created and replicated a 48-item computer administered version of the HDMT (referred to as the Victoria Revision of the HDMT).

Another multiple-choice test implementing the Floor Effect is the TOMM (Tombaugh, 1997). It uses a 2-alternative format to test recognition of 50 line drawings. Using a cutting score of 45 (i.e., 90% correct) on Trial 2, the TOMM correctly classified 95.0% of non-demented patients and 99.9% of cognitively intact volunteers as not feigning, suggesting excellent specificity. Regarding sensitivity, 100% of the simulating participants were correctly identified. Rees, Tombaugh, Gansler, and Moczynski (1998) replicated the above study with similar results.

Rogers et al. (1993) raised a potential ethical concern regarding the administration of tests employing the Floor Effect. For instance, it has been relatively common practice for clinicians to exaggerate the difficulty of easy tasks, such as the FIT, in an effort to increase the test's face validity (Lezak, 1995). However, this practice can be construed as deceiving the client. Rather than resorting to misleading instructions, an alternative may be to provide a cautionary statement that strategies will be employed by the instrument to detect malingered performance (Rogers, 1997).

In summary, the Floor Effect appears to hold promise as a measure of neurocognitive feigning. Its strength lies in its potential for brevity and usefulness as a screen. Its main limitations lie in its vulnerability to detection as a test of malingering and its relatively low probability of producing adequate sensitivity and specificity concurrently. These concerns are more problematic on the simple reproduction tasks, such as the FIT, than on the more complex multiple-choice tasks, such as the VIP

(Frederick & Foster, 1997). Based on Rogers' (1997) classification scheme, the Floor Effect strategy falls within the Probable to Definite range.

Performance Curve

The Performance Curve strategy compares the proportion of easy items correct to the proportion of difficult items correct (Pankratz & Binder, 1997). Performance for genuine patients and controls should decrease as the difficulty of items increases. The underlying assumption is that feigners can be identified because they do not take into account differences in item difficulty when deciding which items to fail. This oversight can result in atypical performance curves (e.g., flat or positively accelerating curves) and can make feigners susceptible to detection. Tenhula and Sweet (1996) collected preliminary data from a debriefing questionnaire that revealed only 16% of simulators took into account the item difficulty when feigning deficits. Thus, the Performance Curve strategy may hold particular promise in detecting malingered performance.

Frederick and Foster (1991) predicted that visual plots of performance for simulators would differ from both cognitively impaired individuals and controls. Specifically, simulators were predicted to have problems keeping track of item difficulty, miss inappropriate items, and exhibit a flat or atypical curve. In contrast, controls and impaired participants were predicted to display negative curves, accurately reflecting their performance. As predicted, potential simulators demonstrated a flat curve, whereas brain-injured and controls demonstrated negative curves.

Instead of relying on visual inspection of PCs, Gudjonsson and Shackleton (1986) identified a single score for PC. They calculated the linear rate of decay across items of

increasing difficulty on Raven's Matrices (Raven, 1958). Using five levels of item difficulty, they found that rate of decay was highly effective at discriminating feigners from genuine patients. Although sample sizes were relatively small, the rate of decay formula yielded a sensitivity rate of 90.0% and a specificity rate of 83.0%.

The PC strategy has been used in other neuropsychological tests as well. For instance, Lezak (1995) suggested that the PC strategy can be used on the Dot Counting Test (DCT) as a measure of feigned memory impairment. Binks, Gouvier, and Waters (1997) found that the DCT provided several different scores that were significantly different when comparing simulators, controls, and patients. The slope of the response latency curve on ungrouped dots was shown to effectively discriminate fakers from non-fakers of neurological dysfunction (hit rate of 85%). The sensitivity rate for feigners was 77.4%; specificity for the same group was 100%.

The utility of the PC strategy when used in a full battery of neuropsychological tests was evaluated by McKinzey, Podd, Krehbiel, Mensch, and Trombka (1997). They investigated the efficacy of a malingering detection formula based on the PC strategy applied to the Luria Nebraska Neuropsychological Battery (LNB). The formula was derived by calculating biserial correlations for malingerers and patients on LNB items and contrasting easy items with more difficult items. Specifically, 34 items that correlated (either negatively or positively) above $\underline{p} < .01$ with group membership (i.e, malingering or patient) were formulated into positive and negative correlation scales. The empirically derived formula appears to contrast performance on simple tasks with performance on more difficult tasks. If more easy items are missed than difficult items, then

"malingering" is suggested. The formula provided a high hit rate of 94.1%, which was maintained upon cross-validation (88.2%).

A potential advantage of the PC and Rate of Decay strategies is that it is difficult for the would-be feigner to continually predict which items should be missed, particularly if the items are not presented in ascending difficulty. Also, PC can be employed on existing neurocognitive measures. A potential disadvantage is that few studies have validated the use of PC and Rate of Decay strategies. The available data suggest that PC and Rate of Decay strategies fall within the Probable Range of classificatory certitude.

Magnitude of Error

The Magnitude of Error (MoE) strategy evaluates the extent to which clients respond incorrectly. The strategy assumes that patterns of near misses and gross errors suggest feigned performance. Much of the existing malingering research has in large part ignored the magnitude of the error and has instead focused simply on whether responses on the feigning test were right or wrong. Martin, Franzen, and Orey (1998) recently used the strategy on modified Visual Reproduction and Logical Memory subtests of the WMS-R (Wechsler, 1987). They designed a multiple choice recognition format for both subtests and asked 10 graduate students in psychology to rate the likelihood of selecting each wrong response. The authors concluded that items chosen by probable malingerers and not by normal controls or brain injured patients (i.e., gross errors) could indicate exaggeration or feigning of deficits. Consistent with their expectations, simulators and suspected malingerers were more likely to select low probability multiple-choice items, and often endorsed choices that very few controls and moderate to severe CHI patients

endorsed. The authors reported classification rates of 86.0% for simulators, 100% for suspected malingerers, 80.0% for actual brain-injured and 100% for controls. A strength of their use of the strategy is that it uses information from commonly-used measures in clinical neuropsychology (visual and auditory memory tests).

The current conceptualization of MoE developed through the Ganser Syndrome literature. As mentioned earlier, Ganser Syndrome is typified by approximate answers; responses that strongly suggest that the patient is somehow aware of the correct response but provides an approximation of it. Bash and Albert (1980) suggested that "near misses" are common in malingering. However, earlier research (Andersen, 1959) indicates that approximate answers are not unique to malingerers but also are found in somatizing patients.

The MoE strategy allows for approximate answers to be identified as qualitatively different from grossly wrong answers. The available data on the MoE strategy yield good classification rates. However, only one study was found that used MoE; thus, conclusions regarding the utility of the MoE strategy in practice should be made cautiously. Until more data are collected, a Tentative classification appears appropriate.

Response Time

The Response Time (RT) strategy assumes that markedly prolonged RT may indicate malingered neurocognitive performance. Only recently has the potential of Response Time (RT) been studied with reference to neurocognitive feigning. To examine the utility of RT as a measure of feigning, Beetar and Williams (1995) compared simulators' and controls' RT on the FIT, Dot Counting Test (DCT), and sections of the

Memory Assessment Scales (MAS; Williams, 1991). Results suggested that simulators indeed took significantly more time to respond to items requiring both recognition and recall. Classification rates could not be calculated due to the lack of a clinical sample; thus, the results are of limited practical value. Nonetheless, the study was one of the first to investigate the utility of the RT strategy on common neuropsychological measures.

Binks et al. (1997) included a clinical comparison group in their examination of the efficacy of RT on the DCT. They measured RT by calculating the total time for ungrouped dot counting minus the total time for grouped dot counting. Although differences in RT were found, they did not discriminate well between all groups. Differences were noted between simulators and neuropsychological patients (with patients taking longer than simulators), but not between simulators and controls. Similarly, Rose et al. (1995) found that RTs were greater in the brain-damaged group than in the simulating group on the PDRT. They suggest that brain-injured participants should be expected to process stimuli slower than unimpaired individuals because of their injury. In contrast, Rees et al. (1998) reported that RTs were higher in simulators than in brain-injured patients on the TOMM (Tombaugh, 1997). They posited that either (a) simulators believe brain damaged individuals take longer to respond, or (b) simulators require more time because they must process both the correct and the incorrect responses. This discrepancy in findings indicates that more research is needed on the applicability of RT in detection of neurocognitive deficits.

Advantages of the RT strategy include: (a) easy application to many existing neurocognitive measures, (b) suitability for computer administration, and (c) easy use in

combination with other detection tests or strategies. A major disadvantage of the strategy is that RT has been shown to be both higher and lower in simulating samples relative to patients. This limitation appears to depend on the particular test chosen for use with RT. Overall, RT appears to fall within the Speculative Range of Rogers (1997) levels of classificatory certitude.

Atypical Presentation

Atypical presentation, or performance that does not make neuropsychological sense, has typically been considered indicative of invalid performance (Lezak, 1995). In other words, when neuropsychological performance is inconsistent with expectations (e.g., Lezak, 1995) or is markedly different across administrations of the same test (e.g., Reitan & Wolfson, 1997) malingering is suspected.

Several studies (e.g., Mittenberg, Azrin, Milsaps, & Heilbronner, 1993; Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner, 1995; Reitan & Wolfson, 1996) have evaluated the utility of Atypical Presentation on specific neuropsychological tests. Reitan and Wolfson (1996) used a differential prevalence design to investigate differences in response consistency over time on the Halstead-Reitan Neuropsychological Battery (HRB) as a means of detecting malingering. In short, they formulated two indices of atypical performance based on within-group and between-group performance on the WAIS-R and HRB. Test-retest performance differences between a litigating group and a non-litigating group of head-injured patients were evaluated. They found significantly more inconsistencies in the litigant than in non-litigant group. Although the results are

intriguing, the assumptions of differential prevalence design (i.e., that litigants are likely to be malingerers) severely limit the utility of the findings.

Mittenberg et al. (1995) examined Atypical Presentation on the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981) in a single testing session. Based on Heaton et al. (1978, p. 492), they defined atypical performance as, "disproportionate reductions in Digit Span relative to other intellectual functions." Because the Vocabulary subtest is thought to be relatively insensitive to brain damage and closely related to overall intelligence, Vocabulary was compared to Digit Span. They found that feigned profiles showed exaggerated relative reductions on Digit Span. A Discriminant Function Analysis (DFA) with the Vocabulary-Digit Span difference as the independent variable accurately classified 70.5% of the participants. While these findings are promising when considered as just one indicator of feigning, the false positive rate is probably too high (36.8%) for clinical use.

Atypical Presentation has also been studied on the California Verbal Learning Test (CVLT, 1987). Sweet et al. (2000) has argued that the CVLT is particularly promising as a measure of feigning because it is a challenging learning and memory task with numerous scores. As a result, patterns of performance can be evaluated for deviations from the expected pattern. The range of task demands and the numerous measures of performance may make it difficult for those attempting to feign to identify how to do so credibly. Trueblood and Schmidt (1993) and Trueblood (1994) found that both the number of correctly recognized words and total number of words recalled

differentiated a group of suspected malingerers from mild TBI patients. Both scores were lower than is typical in normal controls and in severely head injured patients.

Sweet et al. (2000) conducted a replication study on the CVLT using simulators, suspected malingerers, normal controls, and severely brain-injured patients. The authors reported slightly improved classification rates for each group on most of the variables used as measures of feigning. They concluded that Recognition Hits is the optimal measure of feigning on the CVLT. The authors also attempted to take into account base rates by performing a DFA on a more real-world estimate of malingering prevalence. The DFA, with the prevalence rate set at 15.0%, produced a hit rate of 81.1%. Specificities and sensitivities were not published.

Mittenberg et al. (1993) studied Atypical Presentations on indices from the Wechsler Memory Scales - Revised (WMS-R; Wechsler, 1987) to distinguish malingerers from head-injured patients. They found head-injured patients scored relatively better and simulators relatively worse on measures of Attention/Concentration than on other WMS-R scales. The DFA yielded an initial classification rate of 90.5% and cross-validation of 87.2%, suggesting the groups were effectively differentiated. Although these findings are promising, overall interpretation of WMS-R indices remains a controversial topic largely because of deficiencies in the normative sample (McCaffrey et al., 1997). For instance, sample sizes in some cells are very small. Norms for some age ranges, particularly those for older individuals, were extrapolated from other age ranges.

The Atypical Presentation strategy's main strength lies in the fact that no new tests are needed to assess feigning. In addition, potential feigners are not likely aware of

the expected relationships among multiple neuropsychological scores. However, Van Gorp et al. (1999) recently published data to suggest that relying on standard neuropsychological tests to detect feigning is neither sufficiently sensitive or specific. This issue will be addressed further in the discussion chapter. Although there appears to be significant variability in classification rates across studies, most results fall within the Probable Range.

Psychological Sequelae

This technique assumes that unlikely symptom endorsements associated with an injury suggest malingering. Miller and Cartlidge (1972) reported that feigned neurological impairment can often be detected by examining subjective psychiatric complaints. Other researchers (Heaton et al., 1978; Schacter, 1986a) have noted similar increases in reported symptoms in simulators of amnesia.

The utility of Psychological Sequelae as a measure of neurocognitive feigning has been seriously limited by the numerous complaints that are also associated with legitimate head injury. For instance, Lees-Haley and Brown (1993) listed 37 such complaints. Some of these symptoms commonly occur in both normal controls and in brain injury patients. Without corroborative evidence of feigning, the Psychological Sequelae strategy may produce unwarranted classifications of malingering (i.e., inflate the number of false positives).

In summary, only three of the strategies presented have been researched comprehensively: Floor Effect, FCT, and SVT. PC has also been empirically tested but lacks replication; its validation is complicated and it has been operationalized in several

ways. MoE holds considerable promise but also lacks sufficient study. The Atypical Presentation and Psychological Sequelae strategies appear to be the least clinically useful of the strategies. Current operationalizations of these two strategies require further study due to (a) the extensive overlap of symptom endorsement (Psychological Sequelae) and (b) the confounds of unexpected performances of brain-injured patients (Atypical Presentation).

Combined Strategies

Investigators (Nies and Sweet, 1994; Rogers et al., 1993; Rogers, 1997) have recommended that detection of neurocognitive feigning should employ multiple strategies. In line with these recommendations, Frederick and Foster (1991) combined strategies using a multiple-choice modification of the Test of Nonverbal Intelligence (TONI; Brown, Sherbenou, & Johnson, 1982). They used three detection strategies to classify feigners, genuine brain-injured, and controls: performance curve, consistency ratio, and a mathematical product of both. They reported that sensitivity and specificity increased when using a combination of strategies rather than when relying on any single strategy. For example, when using two strategies (i.e., Performance Curve and Consistency Ratio), results yielded good sensitivity but mediocre specificity. However, the slope-consistency ratio interaction as well as a cutting score (i.e., ≥ 83 correct) improved classification to 96.4% of the simulating group and 97.2% of the clinical comparison group. A potential drawback to the study is the small number of individuals in the clinical group ($n = 14$).

An elaboration of the Frederick and Foster (1991) study resulted in the recent publication of the Validity Indicator Profile (VIP; Frederick & Foster, 1997). The authors describe the VIP as an objective measure to evaluate an individual's motivation and effort during neurocognitive testing. Though they proposed a total of six specific indicators, combinations of only three detection strategies were used to document invalid performance: performance curve, multiple measures of response consistency, and a combination of both. Violation of two or more of the indicators of inconsistency indicated an invalid performance.

The VIP consists of both verbal and nonverbal questions designed to assess both "suboptimal effort" and ability. The best utility estimates reported suggest that the nonverbal section of the VIP has a sensitivity rate of 73.5% and a specificity rate of 85.7%. The verbal portion performed less efficiently, demonstrating a sensitivity rate of just 67.3% and a specificity rate of 83.1%. Overall, correct classification was found to be 75.2% for the verbal subtest and 79.6% for the nonverbal subtest. Importantly, it is difficult to interpret the classifications because there are four classification categories (careless, invalid, malingered, valid) rather than the more conventional two categories (feigned vs. honest). Classification rates for feigning per se appear to be much lower (< 10%). The VIP falls between the Tentative and Probable categories of Rogers' classificatory scheme.

Allen et al. (1997) published the Computerized Assessment of Response Bias (CARB) to "quantitatively assess a given patient's attitude (response bias) in taking neuropsychological or neurocognitive tests" (p. 1). The test utilizes the Floor Effect, RT,

and FCT strategies in a multiple-choice format. The authors contend that measures of motivation should not measure ability. Because the CARB correlates only modestly with verbal IQ (r = .26; see Allen et al., 1997) on the Wechsler Adult Intelligence Scale, Revised (WAIS-R; Wechsler, 1981), the authors concluded that the CARB is a measure of motivation rather than ability. However, they admit that the false negative rate (31.2%) is currently too high to rely on the CARB as an independent measure of response bias.

Some disagreement is found in the literature about whether feigning measures should also measure ability. On one hand, positive results obtained from a feigning measure that is insensitive to ability are not confounded by poor cognitive ability. The Floor Effect strategy capitalizes on this by using easy items that do not tap ability level. On the other hand, negative results of feigning provide no further information about the patient. Measures of feigning and ability can provide useful data even of the findings regarding feigning are unremarkable.

In sum, combining strategies appears to have considerable promise though few studies have been conducted. The available classification rates are largely consistent with those of individual strategies. The results from the few studies implementing multiple strategies fall within the Probable Range.

Limitations to Assessment of Neurocognitive Feigning

The preceding review has summarized the utility of many tests of neurocognitive feigning. Despite the extensive research, methodological concerns often limit the interpretability of findings regarding feigning status. Like all experimental research, internal and external validity are often at odds in feigning studies. Either experimental

control is reduced or generalizability declines at the cost of improvements in the other form of validity. Feigning research is most constrained by questionable external validity due to the lack of known-groups of malingerers. All conclusions about malingering are at some point inferred from simulators' or suspected malingerers' performance. Likewise, no gold standard is available to compare the utility of proposed measures. This lack of an appropriate criterion means that there is no way of determining with certainty whether a participant or patient is feigning symptoms independent of the clinical tests or experimental procedures being evaluated (Cercy et al., 1997).

As mentioned, the problems of differential diagnosis and prevalence rates are significant limitations for many types of research, including feigning. In some settings, malingering and the other conditions in question are quite rare and as a result, confident classification is difficult. Additionally, anecdotal evidence suggests that some clinicians may be reluctant to classify individuals as malingerers due to their own inexperience with confronting malingerers and due to the potential cost of being wrong (R. Delaney, personal communication, January 16, 2000). This reluctance may yield under-estimates of the actual prevalence of malingering, which in turn can affect classificatory accuracy when relying on these estimates.

Pankratz and Binder (1997) note that it is common practice to discontinue an evaluation when the results are considered invalid. Thus, whether the performance was due to malingering is often not determined because the assessment is cut short.

Many practitioners do not include tests of neurocognitive feigning in their assessment batteries. Rather, they rely on other data gathered during the assessment to

arouse their suspicions before they administer a feigning measure. In addition, many clinicians miss initial signs of malingering due to inconsistent or inappropriate decision rules, over-confidence in their interpretations of results, and poor standardization techniques (Lees-Haley et al., 1996). Neuropsychologists may avoid tests designed explicitly for feigning detection because they are often considered too time-consuming and potentially lacking in informational value. As a result, these tests are rarely administered as part of a standard battery. A measure of neurocognitive feigning that also provides useful data irrespective of feigning status appears warranted. The Test of Cognitive Abilities (TOCA; Rogers, 1996) was developed in an attempt to address the need.

Preliminary Data on the TOCA

The TOCA (Rogers, 1996) is intended as a measure of cognitive ability and validity of responses. It is a computer-based measure comprised of three sections of multiple-choice questions (see the Method section). Initial validation on a sample of 92 undergraduate students suggests the TOCA may be useful as a measure of neurocognitive feigning (Rogers et al., 1996). For example, simulators scored at or below chance levels (i.e., SVT) significantly more often than honest respondents. Data on patients were not collected; one of the purposes of the current study was to collect data on a clinical sample.

The TOCA's utility as a measure of cognitive abilities (e.g., IQ) was examined in relationship to the Shipley Institute of Living Scale (Shipley, 1946). Overall, the TOCA

correlated at 0.61 with the Shipley. The correlations between specific sections of the Shipley, GPA, and TOCA are found in Table 4.

Performance on the TOCA appears to be adequately generalizable. For instance, scores were not appreciably affected by gender (r_b = .04 to .13) or ethnicity (r_b = .07 to .16). With respect to age, decrements in performance were noted on the Designs section (r = -.33), but not on Sequencing (r = -.08) or Sentences (r = .03). Thus, preliminary data suggest that performance on the TOCA is largely unaffected by demographic variables. Pending further research, the TOCA has promise as measure of feigning and general estimate of cognitive ability.

Table 4

Correlations of the TOCA with Shipley IQ Estimates and GPA

	Shipley			
TOCA	Vocabulary	Abstract	Total	GPA
Sequencing	.28	.52	.54	.24
Designs	.27	.38	.43	.15
Sentences	.56	.41	.57	.25
TOCA total	.40	.54	.61	.26

Note. Preliminary data were collected on a sample of normal controls.

Purposes of the Study and Research Questions

Purposes of the Study

The primary purpose of the study was to determine which strategies or combinations of strategies best differentiate participants asked to simulate brain injury from bona-fide patients and controls. Scales based on the following strategies were developed within the TOCA: (a) Symptom Validity Testing (SVT), (b) Performance Curve (PC), (c) Floor Effect, (d) Magnitude of Error (MoE), and (e) Response Time (RT). In addition, whether or not simulators would modify their feigning when given explicit instructions cautioning them about the presence of detection techniques was evaluated.

The effect of coaching has received attention in the literature; the effect of cautioning has not. Cautioning is similar to coaching in that instructions are provided to the participant prior to testing. This information is intended to influence the participant's response style. Cautioning differs from coaching in that information about how to appear injured is not provided; rather, cautioning informs the participant of the possible ways he or she may be detected. Rogers (1997) argued that the likelihood of dissimulation can be reduced simply listing the specific strategies used to detect feigning. Even if they do decide to feign, they put themselves at risk of detection because they do not know which items pertain to each strategy; they do not know how to alter their responses or how long to take on each item.

Research Questions

The overarching research question to be addressed in the study was based on the thesis that successful feigning is difficult when simulators (SIM) must elude multiple detection strategies simultaneously. Specific detection strategies that were selected on empirical and theoretical grounds were evaluated. The five strategies under investigation were SVT, PC, Floor Effect, MoE, and RT. The following research questions were considered:

Research Question 1. Research Question 1 attempts to answer the question, do groups differ in performance as measured by individual strategies? It is hypothesized that SIM scores on the detection scales would be significantly different from all other groups.

Research Question 2. Research Question 2 is, how well do the strategies classify groups? Classification rates derived from Discriminant Function Analysis [DFA] and cutting scores are used to attempt to answer this question.

Research Question 3. The third research question asks, to what degree does combining strategies improve classificatory utility? For example, it is hypothesized that SIM would be more prone to detection when the Floor Effect and the MoE strategies were used simultaneously. Discriminant Function Analyses are used to assess combined clinical utility.

Research Question 4. The fourth research question addresses the question, does cautioning participants that detection strategies are in place affect feigned performance? It is hypothesized that cautioned simulators' scores on the detection strategy scales (i.e.,

floor effect, SVT, PC, MoE, and RT) will be less likely to be classified as feigning than non-cautioned simulators'.

CHAPTER II

METHOD

Design

The study employed an analogue design. Specifically, a group of simulating malingerers was compared to groups of controls and patients. The chief advantage of this design lies in the level of experimental control and internal validity. Participants without neurological impairment were randomly assigned either to a simulating group (cautioned group or non-cautioned group) or to the control group. The clinical comparison group was comprised of patients with a history of brain injury or cerebrovascular accident (CVA). The inclusion of patients in the design is important because the utility of the measure depends on how well it correctly classifies simulators versus patients. The five groups were (a) cautioned simulators (CS), (b) non-cautioned simulators (NCS), (c) combined group of simulators (SIM), (d) clinical comparison sample (CL), and (e) normal controls (NC). The dependent variables were the strategies employed to detect feigned performance: SVT, Floor Effect, PC, MoE, and RT.

Participants

Consent forms for participation in the study were approved by the Institutional Review Board at UNT (see Appendix A & B). Participants of all five groups were recruited for the experiment if they were free from significant visual and hearing deficits, were familiar with the use of a computer keyboard, and agreed to be tested for approximately one hour. Participants were recruited irrespective of sex or age. A posttest

manipulation check was used to screen out participants who did not understand their instructions or put forth adequate effort. The specific characteristics of the sample, instrument, and procedure are presented in the following sections.

Four non-clinical groups were included in the study: CS, NCS, SIM, and NC. NC consisted of 22 students at the University of North Texas (UNT). The simulating group (SIM) consisted of 60 UNT students who were divided randomly into two groups: CS and NCS. CL participants were referred to outpatient and inpatient services with documented brain damage of at least mild severity.

Only two participants in the clinical group had pending litigation. Both had sustained well-documented severe head injuries and were considered to have a very low likelihood to malinger. Specifically, the severity of their deficits (a) precluded the need to fake and (b) potentially impaired their ability to consciously and consistently alter their performance. Thus, they were included in the group analyses.

Instruments

Screening Measure

Participants in the simulating and control groups were asked to respond to a brief self-report questionnaire regarding their history of closed head injury. A closed head injury was considered any non-penetrating traumatic insult to the head that resulted in an alteration in consciousness. Relevant demographic information (e.g., age, GPA, gender, ethnicity) was also gathered from the screen (See Appendix C). No participant was excluded on the grounds of prior head injury.

TOCA

The TOCA is a computerized measure of response patterns. It employs multiple detection strategies in order to correctly differentiate groups. The TOCA uses a multiple-choice format in which the respondent chooses twice from four possible responses (i.e., each item consists of two trials). There are 112 questions requiring two responses each for a total of 224 responses.

The TOCA consists of three sections: The Sequencing Section (60 questions) assesses the ability to recognize alphanumerical sequences. The Designs Section (22 questions) involves recognition of patterns based on four design parameters: Shape, color, size, and shading. Finally, the Sentences Section (30 questions) assesses verbal comprehension of incomplete sentences.

The TOCA can be configured to employ a warning at the beginning of the test that cautions against faking and includes descriptions of detection strategies. By presenting the respondent with this information, he or she must perform multiple divergent tasks simultaneously to escape detection (e.g., estimating item difficulty while maintaining a plausible reaction time). When present, the cautionary statement appears on the computer screen following the general instructions; it reads as follows:

> WARNING. Every now and then, someone tries to "fake" the test by doing a bad job. Please don't do this. The test has many safeguards to stop persons from faking it. These safeguards identify people who may be faking. Some of the safeguards are listed below:

- The test checks to make sure you miss more of the difficult items than easy items.
- The test checks to see how many seconds you take on each item and compares it to the difficulty of the item.
- The test checks to see if you are trying by comparing your ability on similar items.
- The test checks to see if you made careless errors, particularly on very easy items.
- The test checks to see if you get the same wrong answers as most people do.
- The test checks to see if you make more mistakes than expected by chance (probability) alone.

Don't worry about these safeguards, just put forth your best effort.

Manipulation Check

A manipulation check conducted after the experiment provided valuable information about various performance characteristics of the participants. For example, participants were asked to briefly describe what they were asked to do and whether they understood the instructions. CS, NCS and NC groups were asked to judge the extent to which they were successful and careful at adhering to the instructions and appearing impaired. They were also asked to describe their level of effort and the perceived effect of incentives (see Appendix D).

Procedure for Nonclinical Participants

Recruitment of Participants and Informed Consent

Both simulating groups and the control group were recruited from undergraduates enrolled in Experimental Methods in Psychology Laboratory (course number 3650) at UNT. Test administration was conducted in computer labs in Terrill Hall (Room 220) and GAB (Room 550K) at UNT. Each lab room has at least eighteen personal computers equipped with Pentium processors and Windows NT operating systems. The test was administered in a group setting in which each participant was seated at a table with his or her own computer. The TOCA diskettes were placed in the drives before the students arrived. Participants were randomly assigned to one of three experimental conditions: CS, NCS, or NC. All participants received a consent form apprising them of the study's purpose, anticipated risks and benefits, and extent of confidentiality.

Exclusion of Participants

Participants in CS, NCS, and NC were tested after giving written informed consent. All those recruited for the simulation and control portion of the study were tested. As recommended by Rogers (1997), a manipulation check was performed following the experiment. Nine of the 60 simulating participants were excluded from analysis for not complying with their instructions (see Results section). The remaining 51 simulators and 22 control group participants met all inclusion criteria.

CS Condition

CS received instructions regarding the scenario they were to follow when taking the TOCA and had as much time to prepare for the assignment as they felt appropriate.

Any questions regarding the instructions were addressed during this time. The scenario (adapted from Rose et al., 1998) was kept in sealed, unmarked envelopes, until opened by participants. The scenario is presented below:

> Recently you were in an accident that was not your fault. A car ran a stop sign and hit you. You hit your head on the dashboard and were knocked out for about 15 minutes. Afterwards, you felt "dazed." As a result of the accident you had to stay in the hospital over night with a severe concussion. Because another person caused the accident, you stand to gain a substantial amount of money through a lawsuit against the person responsible for your injuries. In order to receive the financial reward, you must undergo several psychological tests, one of which addresses your concentration, reasoning, and problem solving ability. The purpose of the examination is to determine if there are any long-term effects from your head injury.

The following instructions provide a strategy to employ while taking the TOCA and gave basic information about the effects of head injury:

> Your goal is to take this test in such a way that will convince the examiner that you have suffered permanent brain damage from the accident. You are to produce the most severe problems that you can <u>without</u> making it too obvious to the examiner. That means your "brain damage" must be believable. Major exaggerations, such as remembering absolutely nothing, are easy to detect. If the examiner does not believe that you have any problems you will not win your lawsuit. In addition, you may be fined or jailed for lying.

> Try to remember the following while taking the test: People who have a head injury often have problems paying attention, cannot remember some things that they used to, and do not think as quickly as they did before the injury. Keep these things in mind when taking the test. Remember, try to convince the examiner that you have serious brain damage without being too obvious. You can do this by mimicking the performance of persons who are truly injured.

Next, participants were asked to follow the directions presented on the computer screen and to continue to do so until they finished the test. When prompted by the computer, participants were asked to enter the identification number found at the top right hand corner of the scenario handout.

General guidelines for taking the TOCA are included in the computer program and appear on the screen once it is started (See Appendix E). The cautionary statement was presented on the computer screen to all members of the CS group.[4] Following the caution, specific directions for taking the TOCA and a sample problem are presented. The first item of the TOCA begins on the next screen.

At the end of the first and second sections, the participants were asked to enter their identification numbers at the prompt.[5] At the conclusion of the test, the cursor returns to the C:\ prompt and the program is no longer accessible to the participants. Upon completion of the TOCA, participants were asked to fill out the debriefing questionnaire.

---

[4] In order to seat participants at the appropriate computers, an assistant to the examiner had knowledge of which simulators were warned and which were not (i.e., simulators who received the warning had to be seated at a computer with the warning set to appear at the beginning of the test).

[5] This facilitates administration and analyses of individual TOCA sections.

NCS Condition

The procedure for NCS was identical to that of CS with the exception of the presentation of the caution. Software used by NCS was not configured to include the cautionary statement. However, the scenario, directions, and computer-generated prompts were the same.

NC Condition

The procedure for NC was similar to that of the simulating groups with the exception of the presentation of the scenario. NC received instructions to perform to the best of their ability (See Appendix F). The cautionary statement was included to facilitate future analyses of the effects of such statements on honest respondents.

Incentives

In an effort to increase positive incentive to feign, CS and NCS were informed of a "prize" of $20 awarded to the three most successful "malingerers." NC were encouraged to perform their best and informed that those who obtain the three highest scores receive $20.

All participants were also encouraged to perform according to their instructions as best they can in order to receive all available course credit. In this way, performance was also contingent on individual performance irrespective of other participants' performance. This arrangement may increase motivation to perform as instructed without excessively raising the pressure of competition (Martin et al., 1993).

To increase external validity, a negative incentive to feign was included. All participants in NC, CS, and NCS were informed that the names of the three worst feigners would to be posted in Terrill Hall, the location of the psychology department.

Debriefing

Upon completion, NCS, CS and NC were debriefed (Appendix D). Participants asked to feign were prompted to respond in writing to the questions regarding the scenario. Others responded only to questions not involving feigning. CS, NCS, and NC received proof of extra credit at this time.

Procedure for Clinical Comparison Group

Recruitment of Participants

The CL group consisted of 42 individuals participating in inpatient or outpatient treatment at Plano HealthSouth Rehabilitation Hospital or John Peter Smith – Trinity Springs Pavilion Hospital. Both hospitals receive neuropsychological consultations and referrals from acute inpatient units and from outpatient rehabilitation programs. Trinity Springs also receives outpatient referrals from outside the hospital. Each patient had been referred for a neuropsychological evaluation and were recruited for participation in the study at that time. Participants were tested in the neuropsychologist's office at either location. The examiner remained just outside the office to answer questions if they arose.

Exclusion Criteria

In order to maximize the range of injury severity, the exclusion criteria for CL were liberal. Those who were unable to read and paraphrase the instructions to the TOCA were not included in the study. Two potential participants were excluded on these

grounds. Also, four patients were excluded prior to testing due to severe sensory-motor impairments.

Screening, Assessment, and Debriefing

As a screen of reading ability, participants were asked to read aloud the general instructions for taking the TOCA; all participants could read adequately so no one was excluded due to reading difficulties. The screen for CL included questions regarding their injury. Information regarding location and date of head injury, patient status (inpatient or outpatient), and general demographics was gathered (See Appendix G).

All screenings and evaluations were administered individually by the author or advanced doctoral students in psychology. Prior to testing, participants were given a brief oral description of the study. The participants then provided written informed consent.

The TOCA was administered to one person at a time, with the cautionary statement in place. This arrangement allows for future analyses to investigate the potential effects of warning bona-fide patients about the use of detection strategies. All participants were asked to respond honestly and to perform to the best of their ability.

Once testing had begun, the procedure for testing CL was the same as for the three other groups. Upon completion, the CL group also received a debriefing to ensure any questions regarding the experiment were addressed (See Appendix H).

CHAPTER III

RESULTS

The first main section of this chapter reviews the sample characteristics in the study. Next, the data from the manipulation check and the debriefing are evaluated. The final section contains the performance characteristics of the groups, including means, standard deviations, effect sizes (Cohen's d), F statistics, posthoc comparisons (Tukey's HSD), and utility estimates.

Ethical Considerations of Releasing Test Data

Furnishing information about scale composition or cutting scores of performance on tests of feigning could potentially lead to the coaching of actual malingerers with the data derived from malingering research (see Sweet et al., 2000). As a consequence, specific cutting scores and items used in scales to detect simulators are not included here, but will be made available to qualified researchers upon request.

Sample Characteristics

The following section describes the sample characteristics for the experiment, including the demographic makeup. The 51 simulators and 22 controls included in the study were recruited from UNT Psychology 3650 labs. The 42 clinical participants were recruited from and tested in two local area hospitals (see Chapter II - Method).

Gender was analyzed first and was found to be significantly different among groups, X^2 (2, n = 115) = 6.18, p = .05. While there were more females than males in the

control and simulating groups, the reverse pattern was noted in the clinical group (see Table 5). Of the 115 participants, 64 were female (55.7%). The clinical group was comprised of 40.5% females, while the combined simulating group was made up of 64.7% females, and the control group contained 63.6% females. Performance on the TOCA (Total Score) was not influenced by gender, $t(114) = .23$, $p = .82$.

The ethnic composition of the sample was also examined. The majority of the participants were Caucasian (85.2%); the remaining 14.8% was comprised of Asian Americans (1.0%), African Americans (1.0%), and Hispanic Americans (12.6%). An ANOVA could not be performed on the entire sample due to sample limitations. A t-test performed on the two largest subgroups (Caucasians and Hispanic Americans) failed to find significant differences in Total Score on the TOCA, $t(106) = 1.98$, $p = .06$. Whether performance on the detection scales varied by ethnicity was examined in a supplementary analysis and is discussed in Appendix K.

Examination of level of education revealed that all participants achieved at least a high school education and four participants (all CL) had taken graduate level courses. The vast majority (90.5%) of participants had taken sophomore level or higher college courses. All non-clinical participants ($n = 73$) had taken "some college classes," while 69.2% of CL reported having taken some college classes. About 12% of CL had not graduated from high school. Level of education did not significantly affect Total Score on the TOCA, $F(3,106) = 1.33$, $p = .27$.

The next analysis involved the age of the participants. Age varied significantly by group (see Table 6). The average age of the participants in the study was 31.9 years (SD

Table 5

Participants' Gender by Experimental Condition.

Gender	NC	SIM	CL	Total	X^2	df	p
Male	8	18	25	51	6.18	2	.05
Female	14	33	17	64			

Note. For Groups, NC = Normal controls, SIM = Both simulating groups, and CL = Clinical participants.

Table 6

Age of All Participants and GPA for NC and SIM.

Age	Mean	SD	Min	Max
NC	24.68_a	6.16	19	43
SIM	25.87_a	7.10	17	46
CL	43.80_b	17.36	17	84
Total	31.58	14.26	17	84
GPA[a]				
NC	3.28	0.48	2.4	4.0
SIM	3.17	0.51	1.8	4.0
Total	3.20	0.51	1.8	4.0

Note. For age, $F(2, 114) = 33.04$, $p < .001$. Means with different subscripts differ at $p < .01$ by the Tukey test. For GPA, $t(115) = .35$, $p = .70$. For Groups, NC = Normal controls, SIM = Both simulating groups, and CL = Clinical participants.
[a] GPA was not collected for CL.

= 14.49). The control and the simulating groups were approximately the same age, with mean ages of 25.1 and 25.5, respectively. Members of the clinical group were older on average, with a mean age of 43.4. Age correlated with Response Time (RT) at .41. The implications of this correlation are discussed in the section on the RT detection strategy.

GPA was analyzed among non-clinical groups only and did not differ among these groups, F(2, 77) = .354, p = .70. The mean GPA for the sample was 3.20 (SD = 0.50) and ranged from 1.80 to 4.00 (see Table 6). Among non-clinical participants, GPA was significantly correlated with Total Score, r = .27, p = .04. This finding is consistent with the pilot data (r = .26).

Finally, location of elementary school attended (within or outside the U.S.) was obtained. The rationale for obtaining this data came from the fact that Section 3 of the TOCA requires knowledge of common American children's stories. Nearly all participants (92.6%) received their education within the U.S. only, while very few (3.7%) were educated both within and outside the U.S., or outside of the U.S. exclusively (3.7%). All participants learned English as their first language.

Head Injury and Diagnostic Data from Non-clinical Groups. This section examines the effects of pre-existing conditions, such as a brain injury or psychiatric illness on the TOCA. These data were considered important due to their potential effect on current level of neurocognitive functioning. It is important to determine both the frequency of the premorbid conditions in the present sample and the degree to which they may have influenced performance on the TOCA. Of the 73 non-clinical participants, 62 provided data regarding premorbid conditions. Nine (14.5%) reported having experienced a closed head injury with a loss of consciousness (LOC). Seven of the 9 (77.8%) reported LOC of several minutes or less and 2 participants (22.2%) reported being unconscious for at least one hour. Five (55.6%) required a visit to the ER; however, only one (11.1%) required an overnight stay. Both individuals who reported a loss of consciousness greater

than one hour also reported that the injuries occurred when they were children. Importantly, both have obtained college degrees and denied having difficulties in school. They also did not report any persisting sequelae from the head injuries. Nevertheless, the performances of these two individuals were analyzed separately from the other less seriously head-injured participants. One of the two participants scored above the mean on each section of the TOCA. The other participant scored below the mean on Sections 1 and 3 but slightly above the mean on Section 2 (see Table 7). All scores for each participant were within one standard deviation of the mean; thus, they were retained for analysis.

The other seven simulators with self-reported head injuries did not differ in performance from that of the other simulators and normal controls (Section 1, $\underline{t}(71) = 1.43$, $\underline{p} = .16$; Section 2, $\underline{t}(71) = 1.33$, $\underline{p} = .18$; Section 3, $\underline{t}(71) = .26$, $\underline{p} = .79$. Nor did they differ in mean GPA, $\underline{t}(71) = .72$, $\underline{p} = .47$. As a consequence, these cases of brief LOC were retained for subsequent analyses.

As mentioned, mental disorders can influence current neurocognitive functioning and confound the results of analyses. Self-reported histories of Learning Disorder (LD), Attention Deficit-Hyperactivity Disorder (ADHD), anxiety, and depression were obtained in the demographics questionnaire.[6] Of the 115 entries, 13.0% endorsed having LD, 8.7% reported ADHD, and 0.9% reported depression or anxiety. Most participants (77.4%) reported that they had never received a diagnosis of LD, ADHD, anxiety disorder, or

[6] These conditions were considered valid if the participant reported being evaluated and diagnosed by a mental health professional. Other reports (e.g., "my mother/friend/spouse said I was hyperactive") were excluded.

Table 7

TOCA Performance of Two Simulators who Reported Loss of Consciousness for More Than One Hour Compared to Performance of other Simulators.

	Score for		Mean (and SD) for
	Subject #117	Subject #149	Other Simulators
Section 1	124	164	152.55 (41.72)
Section 2	34	38	33.10 (10.67)
Section 3	78	102	83.25 (23.98)

depression. The 22.6% of the participants with a prior mental disorder did not differ from the other simulators and controls on TOCA performance, Section 1, $F(2, 71) = 3.66$, $p = .10$, Section 2, $F(2, 71) = .35$, $p = .56$, Section 3 = $F(2, 71) = .15$, $p = .69$. Because a trend was noted in Section 1, a supplementary analysis was performed to evaluate possible differences in likelihood of being detected as feigners based on history of a mental disorder (see Appendix L). No differences in GPA were noted between those with a history of psychiatric disorder and the other participants, $t(71) = .63$, $p = .53$. Given these data and the participants' academic achievement (e.g., taking college courses), these cases were included in the study.

Manipulation Check

Participants of both simulating groups were asked to describe their instructions for the experiment. As a manipulation check, 8 SIM (9.7%; 2 from NCS and 6 from the CS groups) were excluded because they mistakenly described the experimental condition (i.e., they reported that they were to perform the best they could). Given that they were actually instructed to feign a head injury, this finding suggests that they did not likely understand their role described by the instruction set. CS in particular may have answered

to the best of their abilities because the caution at the beginning of the TOCA recommends that they do so. NCS did not receive such a caution. As a result, CS may have become confused as to whether they should feign or they should "just do your best job," as instructed by the caution. The eight SIM were excluded from subsequent analyses.

Simulators were asked how carefully they adhered to the instructions to feign (see Table 8). All participants said that they had been at least "somewhat careful" to perform according the instructions, and 57.2% said that they were "quite careful to very careful." Approximately 81% of SIM rated themselves as at least "a little successful" at playing their role according to the instructions. Roughly 40% said that they were "quite" or "very" successful. The remaining 19.1% said they were "not successful." Perceived success did not significantly affect performance on the TOCA (Total Score), $\underline{F}(2, 46) = 2.22$, $\underline{p} = .31$.

The degree of effort put forth during the experiment can potentially influence the results and subsequent interpretation of low scores. Over 91% of the participants reported that they gave good or great effort, while 8.4% said that they gave "a little effort" (see Table 8). Total Score on the TOCA did not vary according to level of perceived effort, $\underline{F}(2, 68) = 1.23$, $\underline{p} = .30$. Regarding their responses to incentives, 58% of the participants endorsed trying at least "a little harder" to obtain (or avoid, for negative incentive) incentives, while 42% did not try harder due to incentives. Perceived response to incentive did not significantly affect performance on the TOCA, $\underline{F}(4, 51) = .81$ $\underline{p} = .53$.

Table 8

Simulators Perceived Success, Carefulness, and Effort

	Perceived Success, Carefulness, and Effort (%)			
Variable	Not at all	A Little/Somewhat	Quite/Good	Very/Great
Carefulness	0	42.8	35.8	21.4
Effort	0	8.4	81.4	10.2
Success	19.1	40.6	36.5	3.8

Note. Participants (n = 8) inaccurately reporting the instructions for their experimental condition were excluded from this table.

The manipulation check also examined whether participants tended to use particular strategies to appear brain-injured. About 24% of the simulators said they used only one strategy to appear brain-injured, (10.9% of SIM said they answered wrong on purpose, 6.5% said they gave inconsistent effort, and 6.5% slowed their RT); 71.7% of respondents said they used a combination of all three of the above strategies. Responses regarding what they did during preparation time suggest that most read and reviewed the instructions (62.5%), while fewer participants used the time to consider possible strategies for faking convincingly (10.0%).

All simulators were asked what may have hindered their ability to fake convincingly, irrespective of perceived success, effort, or carefulness. Of three possible choices, the most common response was that it was “too hard to fake on this test” (53.6%); followed by, “I am too honest to fake” (25.8%), and “more than one factor made it difficult to fake” (19.1%). One individual, who had a learning disability, reported

that she was unable to understand the directions and was excluded from subsequent analysis.[7]

Another purpose of the manipulation check was to determine how many participants perceived themselves as successful at feigning. Of those that felt they were at least "a little successful" at faking, almost 20.0% said their knowledge of brain function helped them, while another 20.0% said they were good at deceiving people. A variety of other explanations for what helped SIM fake was offered. For example, some participants mentioned that they focused on the advice in the scenario; others considered it a challenge to fake convincingly; and still others said it was a combination of these factors. One member of the SIM group noted that malingering was the topic of a recent lecture. This category of heterogeneous factors that helped them fake made up approximately 60.0% of the responses (see Appendix I).

Additional Data from Clinical Group

Table 9 contains information regarding patient status, employment, injury severity, and diagnosis. Outpatients represented the majority of patients, followed by persons living in the community and inpatients. Most CL were employed part or full-time prior to their injuries, while less than 10% were students. Analyses at the end of this chapter summarize the potential effects of injury severity and time since injury on the TOCA's ability to differentiate feigned from honest performance.

Roughly one-third of the patients sustained primarily left cerebral hemisphere damage, while almost half of the sample sustained diffuse damage (see Table 9). The

[7] This individual's exclusion increased the total excluded to nine.

remaining patients suffered primarily right hemisphere injuries. Presence of lateralized lesions versus diffuse damage was determined by radiography, neuropsychological test results, and other lateralizing signs (e.g., hemiparesis and aphasia).

The time since the injury was also evaluated. On average, patients were 36.41 months post injury; however, times were highly variable (SD = 62.89). Time since injury was divided into recent (< 6 months) and older (> 6 months) injuries in order to briefly assess the effect of time since injury on TOCA performance. No difference in performance was observed, $t(41) = 1.11$, $p = .33$.

Table 9

CL Sample Characteristics: Patient Status, Employment Status, Injury Severity and Location, and Diagnosis

Variable	N	(%)
Patient Status		
Inpatient	5	11.9
Outpatient	30	71.4
Community	7	16.7
Employment History		
Unemployed	10	24.3
Employed	28	65.9
Student	4	9.8
Injury Severity[a]		
Mild	25	59.5
Severe	15	40.5
Injury Location		
Diffuse	20	47.6
Right	9	21.5
Left	13	30.9
Diagnosis		
CVA	19	45.3
TBI	23	54.7

Note. TBI = Traumatic brain injury; CVA = Cerebrovascular accident.
[a] Data regarding location of cerebral injury were unavailable for two cases.

In terms of diagnosis, just over one-half of the patients received a diagnosis of traumatic brain injury (TBI) and the remaining patients were diagnosed with a cerebro-vascular accident (CVA; see Table 9). Diagnoses were determined by a review of medical records for each patient.

Performance Characteristics

ANOVAs were conducted to analyze the effect of group membership (i.e., experimental condition) on performance on each section of the TOCA. Section Score was used as the DV and Group (NC, CS, NCS, and CL) as the IV. Only 1 of the 4 groups (NC) was both neurologically intact and instructed to perform to the best of their ability; thus, it was predicted that a significant difference would be found between NC and (a) CS, (b) NCS and (c) CL. The results revealed that group performance differed significantly on all sections. Tukey's honestly significant difference (HSD) comparison revealed that NC did perform significantly better than all other groups on Section 1 (see Table 10). On Section 2, Tukey's HSD revealed a significant difference between CL and NC. On Section 3, the CL and NC groups scored significantly better than both simulating groups, but did not differ from each other.

Analyses of Research Questions

The detection strategies used in the study are operationalized in Table 11. Each strategy will be addressed individually, as stipulated by Research Question 1. Information regarding scale composition, item/response derivation, and operational definitions can be found at the beginning of each strategy's results section. The first section analyzes group differences by strategy (Research Question 1). Next, the utility of using the strategies

Table 10

Group Means Comparisons by Section Score with F Statistics and Tukey HSD (SDs in Parentheses)

	Group				
	Honest	Simulators			
Variable	CL	NC	CS	NCS	F
Section 1	176.29_a (56.94)	205.64_b (27.81)	149.09_a (37.04)	155.17_a (45.41)	6.13^c
Section 2	27.29_a (10.05)	36.64_b (12.35)	34.27_{ab} (12.12)	32.21_{ab} (9.53)	3.44^d
Section 3	103.09_a (16.07)	109.27_a (7.72)	84.00_b (21.62)	82.69_b (26.23)	13.26^c

Note. Group means with different subscripts are significantly different, Tukey's HSD, $p < .05$. For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls.

[c] $p < .001$

[d] $p < .01$

independently, then in combination, is addressed (Research Questions 2 and 3). Lastly, the effect of cautioning simulators that detection strategies are used is presented (Research Question 4).

Research Question 1

The first research question examined the utility of each feigning detection strategy in revealing differences in performance among groups (NC, CS, NCS, and CL). Each of the five strategies was analyzed independently by an ANOVA. Means, F statistics, and an

Table 11

Operational Definitions of Detection Strategies Employed by the TOCA

Floor Effect	
FE-90.0%-I	Based on comparison of individual performance on the 13 items answered correctly by at least 90.0% of the clinical comparison group.
FE-95.0%-I	Based on comparison of individual performance on the 7 items answered correctly by at least 95.0% of the clinical comparison group.
FE-95.0%-R	Based on comparison of individual performance on the 22 responses answered correctly by at least 95.0% of the clinical comparison group.
Magnitude of Error	
7.1%-MoE1	It is a comparison of performance on gross errors. Participants are classified as feigning if they choose incorrect responses chosen by less than 7.1% of CL in Section 1.
7.1%-MoE3	It is a comparison of performance on gross errors. Participants are classified as feigning if they choose incorrect responses chosen by less than 7.1% of CL in Section 3.
Truncated	It is a comparison of performance on gross errors. This scale consists of MoE items of medium difficulty only (items answered correctly by 25 – 50% of CL). Participants are classified as feigning if they choose incorrect responses chosen by less than 10% of CL.
Response Time	It is a comparison of latencies from stimulus presentation to response, measured in seconds for each section (RT1, RT2, RT3, & RT Total).
Performance Curve	(a) It is a visual comparison of individual performance curves as item difficulty increases. Item difficulty is divided into quartiles. (b) Ratios of difference scores between items of varying difficulty are compared.
Rate of Decay	Based on the formula (2A + B) – (D + 2E) reported by Gudjonsson & Shackleton (1986). Item difficulty is divided into quartiles.
Symptom Validity Testing	
SVT-1st	Based on chance probabilities, participants are classified as feigning if they fail to answer correctly the first answer to each question at least 25.0% of the time.
SVT-Either	Based on chance probabilities, participants are classified as feigning if they fail to answer correctly either of the responses of each item at least 41.0% of the time.
SVT-Both	Based on chance probabilities, participants are classified as feigning if they fail to answer correctly both parts of each question at least 8.0% of the time.

estimate of effect size (Cohen's d) for each strategy are found in Tables 13 - 21. All Fs were significant and ds ranged from 0.11 to 2.20.

Floor Effect. The Floor Effect strategy has received a great deal attention in the literature and has been shown to be an effective means for detecting feigned neurocognitive impairment. Briefly, it compares the number of very easy items correct to previously established performance standards. Three scales based on the Floor Effect strategy were established for this study: FE-90% Items, FE-95% Items, and FE-95% Responses. Table 12 displays items and responses that comprise the three Floor Effect scales.

- FE-90%-Items Scale. Items answered correctly by at least 90% of CL comprised the first Floor Effect scale. The 90% criterion resulted in 13 items. An ANOVA performed on these items revealed significantly different performances among groups, $F(3, 111) = 11.90$, $p < .01$. Tukey's HSD indicated that CL differed from both simulating groups ($ps < .01$) but not from NC ($p = .81$). The FE-90%-Items scale yielded the largest effect size of the three strategies. Table 13 contains a summary of the results for each of the Floor Effect scales.
- FE-95%-Items Scale. As a more stringent test, items answered correctly by at least 95% of CL were used. Seven items met this criterion. An ANOVA on the items revealed that performance differed significantly by group, $F(3, 111) = 11.44$, $p < .01$. Tukey's HSD revealed that CL again differed from simulators ($ps < .01$) but not from NC ($p = .96$).

Table 12

Items Employed in Three Floor Effect Scales

Scale	Number of Items	Items
FE-95%-Items	7	5, 93, 102, 103, 108, 110, 115
FE-90%-Items	13	5, 18, 43, 93, 102, 103, 104, 105, 107, 108, 110, 115, 116
FE-95%-Responses	22	2, 9, 10, 35, 85, 185, 186, 203, 204, 205, 206, 208, 214, 215, 216, 219, 220, 229, 230, 231, 232, 236

Table 13

Group Mean and Effect Size (Cohen's d) Comparisons for Three Floor Effect Scales

	Group					
	Honest		SIM			
Floor Effect	CL	NC	CS	NCS	F	Cohen's d
FE-90%-Items (13 items)	12.21_a (1.87)	12.77_a (0.53)	9.36_b (3.27)	9.37_b (3.87)	11.90[d]	1.22
FE-95%-Items (7 items)	6.71_a (0.80)	6.86_a (0.35)	5.36_b (1.79)	5.27_b (1.68)	11.87[d]	1.13
FE-95%-Response (22 responses)	21.29_a (1.56)	21.68_a (0.72)	18.09_b (4.36)	17.48_b (5.39)	10.73[d]	1.08

Notes. For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls. Group means with different subscripts are significantly different by Tukey comparison, $p < .01$. Given the lack of difference between CS and NCS scores, the effect size measures the magnitude of difference between mean CL score and mean SIM score. $d = M_1 - M_2$ / pooled SD.
[d] For F ratios, $p < .01$.

- FE-95%-Responses Scale. Items on the TOCA consist of two responses each. FE was used to evaluate both responses to each item. Twenty-two responses were answered correctly by at least 95%. Significant differences were noted

on these items among groups, $F(3, 111) = 10.73$, $p < .01$. Tukey's HSD revealed significant differences between CL and simulators; no difference was noted between CL and NC ($p = .97$; see Table 14).

Magnitude of Error (MoE). The MoE strategy evaluates the degree to which a response is incorrect. Although opposite ends of the frequency distribution are examined, the rationales for MoE and Floor Effect strategies are similar. Both strategies assume that items that are highly likely to be answered consistently in a given direction can differentiate groups. In contrast to the Floor Effect strategy, the MoE strategy uses the frequency of highly unlikely errors, rather than the frequency of easy questions answered correctly.

The MoE strategy involved an individual response analysis on each section (see Table 14 for scale composition); more frequent endorsement of unlikely response alternatives in SIM than in CL was considered evidence of neurocognitive feigning.

Two empirical approaches, one established a priori and one post-hoc, were utilized. The a priori approach was used to limit items to those of medium difficulty. Of these items, only those answered incorrectly by less than 10% of CL were used. This truncated approach removed items that were unlikely to contribute to the discriminatory power of the scale (i.e., those at the extremes).

Table 14

Magnitude of Error (MoE) Scale Composition for each of the MoE Scales

MoE Strategy	Responses
7.1%-MoE1 Rule	3, 5, 6, 7, 9, 10, 12, 13, 14, 16, 17, 18, 19, 25, 26, 28, 29, 30, 31, 32, 35, 36, 53, 63, 65, 79, 81, 82, 84, 85, 86, 92, 95, 96, 98, 105, 106, 110
7.1%-MoE3 Rule	186, 189, 190, 191, 192, 197, 198, 202, 203, 204, 205, 206, 207, 208, 209, 210, 213, 214, 215, 216, 219, 220, 221, 222, 231, 232, 234, 236
Truncated MoE (10%)	1, 5, 15, 16, 17, 19, 20, 21, 22, 23, 24, 29, 30, 37, 38, 41, 42, 47, 48, 51, 52, 53, 54, 55, 58, 73, 74, 75, 76, 80, 81, 82, 83, 84, 89, 90, 99, 100, 101, 102, 109, 113, 114, 115, 116, 119, 120, 125, 126, 181, 193, 195, 219, 220, 225, 226, 227, 228, 237, 238

Note. Items of the 7.1% MoE scales were drawn from Sections 1 and 3 only.

The post hoc approach drew from incorrect responses that were chosen frequently by SIM but seldomly by CL (7.1% or less).[8] Through post-hoc inspection of individual responses, two scales were developed: one from items in Section 1, and one from items in Section 3. Only a small portion of items were answered correctly on Section 2 (3 of 22 items were answered correctly by 50% or more of NC, and only 1 was answered correctly by 75%). This precluded the development of a MoE scale based on performance on Section 2.

[8] 7.1% = 3 of 42 clinical participants. Upon visual inspection of the raw data, three was found to be the optimal cutting score for discriminating SIM from CL.

- 7.1%-MoE1 Scale. In Section 1, 38 incorrect responses were identified that 3 or fewer CL and 4 or more SIM had chosen. An ANOVA revealed that the groups differed significantly in performance, $F(3, 111) = 19.89$, $p < .01$. Tukey's HSD revealed that both simulating groups differed from both honest groups (CL and NC), but did not differ from each other (see Table 15).
- 7.1%-MoE3 Scale. The "7.1% or less" rule was applied to Section 3 in the same way as described for Section 1. Twenty-eight responses were found to meet criteria. When subjected to ANOVA, significant differences were noted among groups, $F(3, 111) = 37.68$, $p < .01$. Tukey's HSD again identified differences between SIM and CL as well as between SIM and NC. This version of the MoE strategy yielded the strongest effect size (Cohen's $d = 1.53$; see Table 15). Combining the two 7.1%-MoE scales revealed the same pattern of findings, $F(3, 111) = 33.39$, $p < .01$.
- Truncated MoE Scale. A second MoE approach was established by an a priori decision to truncate the number of items that could be included: Only those items answered incorrectly by 25 – 50% of CL (i.e., items of medium difficulty). Items were drawn from throughout the TOCA. The rationale for the truncated MoE approach was to eliminate items that very few would miss and those very difficult items that would likely promote guessing. Both very easy and very difficult items may diminish the effectiveness of MoE. Of these items, incorrect responses chosen by less than or equal to 10% of CL were used. Thirty-two responses met these criteria. Group performance differed

significantly on the Truncated MoE scale, $F(3, 111) = 5.02$, $p < .01$ (see Table 15).

Response Time (RT). RT was recorded in seconds automatically by the computer. Four measures of RT were utilized: The time to complete each of the three sections and overall completion time (RT1, RT2, RT3, RT-Total). Age was used as a covariate in analyses of RT for two reasons: (a) RT scores for each group were found to be significantly correlated with age (mean $r = 0.41$, $p < .01$) and (b) members of CL were significantly older than the other groups, $F(3, 111) = 20.97$, $p < .01$. The results of the ANCOVA revealed significant differences for each measure of RT across groups (see Table 16). Specifically, CL took far longer, on average, than SIM and NC to complete the test. Supplemental analyses on RT are discussed at the end of this chapter.

Performance Curve (PC). The PC strategy was designed to reveal individual differences in the slope across items of increasing difficulty. SIM were thought to be vulnerable to detection because they must decide which items to miss without knowing how difficult the items will get. As a result, they may produce flat or atypical performance curves. In contrast, respondents responding honestly provide negatively accelerating curves as items increase in difficulty. Four groups of items based on their difficulty, were drawn from the TOCA: Very Easy, Moderately Easy, Moderately Difficult, and Very Difficult.

All groups' PCs were compared to the control group's PC. The rationale for this was that normal controls were the only neuropsychologically intact participants who were asked to perform their best. Thus, their PCs should reflect normal curves. Certain

Table 15

Group Mean and Effect Size (Cohen's d) Comparisons with F statistic for Four Magnitude of Error Scales

Magnitude of Error	Group: Honest CL	Honest NC	SIM CS	SIM NCS	F	d
7.1%-MoE1	7.14_a (8.32)	3.27_a (2.28)	21.36_b (14.07)	17.59_b (14.07)	15.89[d]	1.04
7.1%-MoE3	1.19_a (2.76)	0.36_a (0.70)	11.09_b (9.49)	9.17_b (9.32)	18.68[d]	1.53
MoE1 & MoE3 Combined	8.33_a (10.12)	3.64_a (2.34)	32.45_b (20.94)	26.76_b (20.99)	21.39[d]	1.32
MoE Truncated	1.59_a (2.74)	0.68_a (0.72)	2.65_b (1.65)	3.16_b (2.23)	5.02[e]	0.35

Note. For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls. Given the lack of difference between CS and NCS scores, the effect size measures the magnitude of difference between mean CL score and mean SIM score. $d = M_1 - M_2$ / pooled SD.

[d] For F ratios, $p < .001$

[e] For F ratio, $p < .01$

Table 16

Group Mean and Effect Size (Cohen's d) Comparisons with F statistic for Four RT Scales

	Group					
	Honest		SIM			
Response Time [f]	CL	NC	CS	NCS	F[e]	d
RT1	40.03_{a} (19.32)	26.21_{b} (6.27)	20.56_{b} (8.13)	21.63_{b} (6.21)	9.54[d]	1.41
RT2	19.06_{a} (4.91)	16.46_{ac} (3.91)	14.38_{bc} (4.36)	13.95_{b} (5.09)	4.75[d]	0.96
RT3	16.29_{a} (10.22)	9.53_{b} (3.22)	13.70_{b} (4.73)	12.07_{b} (3.78)	3.13[d]	0.49
RT-Total	79.18_{a} (27.50)	51.85_{b} (9.32)	48.64_{b} (14.06)	47.65_{b} (12.15)	8.80[d]	1.33

Notes. For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls. Given the lack of difference between CS and NCS scores, the effect size measures the magnitude of difference between mean CL score and mean SIM score. $d = M_1 - M_2$ / pooled SD.

[d] For F ratios, $p < .01$.

[e] ANCOVA with age as covariate.

[f] Average time to complete each section, in minutes.

deviations from this standard curve were considered atypical and indicative of feigning. Items answered correctly by at least 95% of NC were considered Very Easy. Moderately Easy items were those answered correctly by 81.8% - 90.9% of NC. Moderately Difficult items were answered correctly by 54.5% – 77.3% of NC. Very Difficult items were answered correctly by no more than 50.1% of NC. These ranges were selected in an attempt to separate levels of difficulty as clearly as possible while maintaining an equal number of items in each level part. They also reflect idiosyncrasies of the data distribution that restricted the number of alternative levels of difficulty. Table 17 contains

Table 17

Group Mean and Effect Size (Cohen's d) Comparisons with F statistic for the Performance Curve Strategy

Level of Item Difficulty	Group: Honest CL	Honest NC	SIM CS	SIM NCS	F	d
Very Easy	93.33_{a} (11.55)	102.73_{a} (1.14)	76.36_{b} (17.27)	79.86_{b} (20.90)	11.95[d]	0.99
Moderately Easy	81.48_{a} (19.89)	93.91_{b} (4.73)	67.64_{ac} (16.03)	68.41_{ac} (20.35)	10.86[d]	0.71
Moderate Difficulty	68.81_{ab} (16.33)	82.18_{a} (13.74)	58.18_{b} (12.14)	62.62_{b} (15.04)	6.20[d]	0.56
Very Difficult	38.24_{a} (10.18)	45.45_{b} (12.28)	43.00_{ab} (10.04)	38.41_{ab} (9.38)	2.41[e]	0.23

Notes. For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls. Given the lack of difference between CS and NCS scores, the effect size measures the magnitude of difference between mean CL score and mean SIM score. $D = M_1 - M_2$ / pooled SD.
[d] For F ratios, $p < .01$.
[e] For F ratio, $p < .05$.

the means and SDs at each level of difficulty. Table 18 displays the items that comprise each level of difficulty. Within-subjects analyses via MANOVA on Section 1 revealed that performance was significantly different within each group across item difficulty, $F(3, 111) = 83.78$, $p < .01$, suggesting that items were indeed differentially difficult. Between-group differences can be found in Table 19.

Figure 1 displays performance curves for CS, NCS, CL, and NC based on performance on all 3 sections of the TOCA. Group means are plotted by level of

Table 18

Performance Curve Items from All Sections by Level of Difficulty and by Percent Correct in NC

Descriptor	Items	Percent Correct
Very Easy	2, 5, 7, 9, 10, 13, 14, 15, 17, 18, 33, 43, 92, 95, 101, 102, 103, 104, 105, 107, 108, 110, 111, 115, 116, 117	≥ 95.5
Moderately Easy	4, 6, 8, 11, 12, 16, 19, 20, 27, 28, 31 37, 40, 42, 49, 50, 55, 58, 91, 93, 97, 99, 100, 106, 112, 120	81.8 - 90.9
Moderately Difficult	1, 23, 24, 25, 26, 29, 30, 32, 34, 35, 36, 38, 39, 41, 46, 47, 51, 54, 56, 57, 63, 87, 98, 113, 114, 119	54.5 - 77.3
Very Difficult	21, 44, 45, 52, 59, 60, 61, 62, 64, 67, 69, 70, 72, 74, 75, 77, 78, 79, 81, 82, 84, 85, 88, 89, 90, 94	≤ 50.1

Note. There are 26 items per level of difficulty.

difficulty, with item difficulty increasing left to right. Contrary to the hypothesis, the figure reveals no grossly atypical curves; rather, each group produced a negatively accelerating curve.

The PC strategy was also applied to individual sections of the TOCA. Visual inspection of Figure 2 reveals that, consistent with the overall performance curves (Figure 1), scores were higher for NC than for the other three groups at all levels of item difficulty in Section 1. Both simulating groups performed worse than NC and CL at all

levels of difficulty. In contrast to the overall curves (Figure 1), CS displayed an atypical curve, missing more Moderately Easy and Moderately Difficult items than Very Difficult items.

Table 19

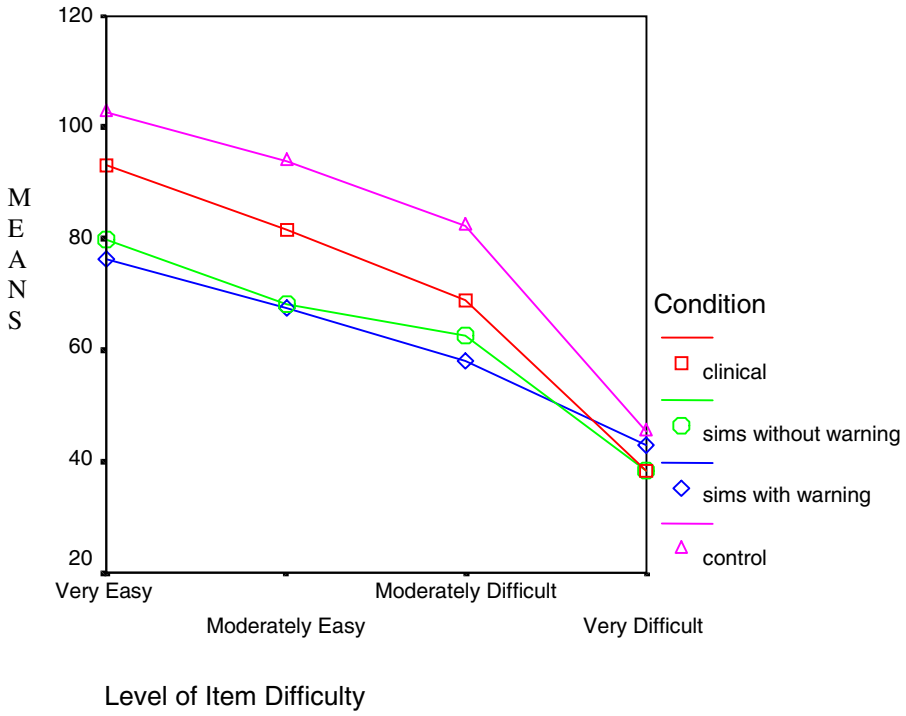
Group Mean and Effect Size (Cohen's d) Comparisons for each Level of Item Difficulty (Performance Curve) on Section 1

	Group				
	SIM		Honest		
Level of Difficulty	CS	NCS	CL	NC	F
Very Easy	31.00_{ab} (7.95)	30.00_{a} (9.25)	33.24_{ab} (10.85)	39.18_{b} (2.20)	5.10^{c}
Moderately Easy	24.82_{a} (7.35)	27.66_{a} (8.72)	30.29_{ab} (11.16)	36.09_{b} (4.47)	6.56^{d}
Moderately Difficult	22.18_{a} (8.94)	24.83_{a} (10.10)	26.62_{a} (12.34)	33.73_{b} (7.26)	5.10^{c}
Very Difficult	24.82_{ab} (7.11)	24.41_{a} (8.18)	26.62_{ab} (11.40)	32.00_{b} (7.33)	3.34^{c}

Notes. Means with different subscripts are significantly different by Tukey comparison, $p < .05$. For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls.
Cohen's d was not calculated due to the lack of statistical difference between CL and SIM.
[c] $p < .05$
[d] $p < .01$

Figure 1

Performance Curves for CS, NCS, NC, and CL on TOCA (all sections)

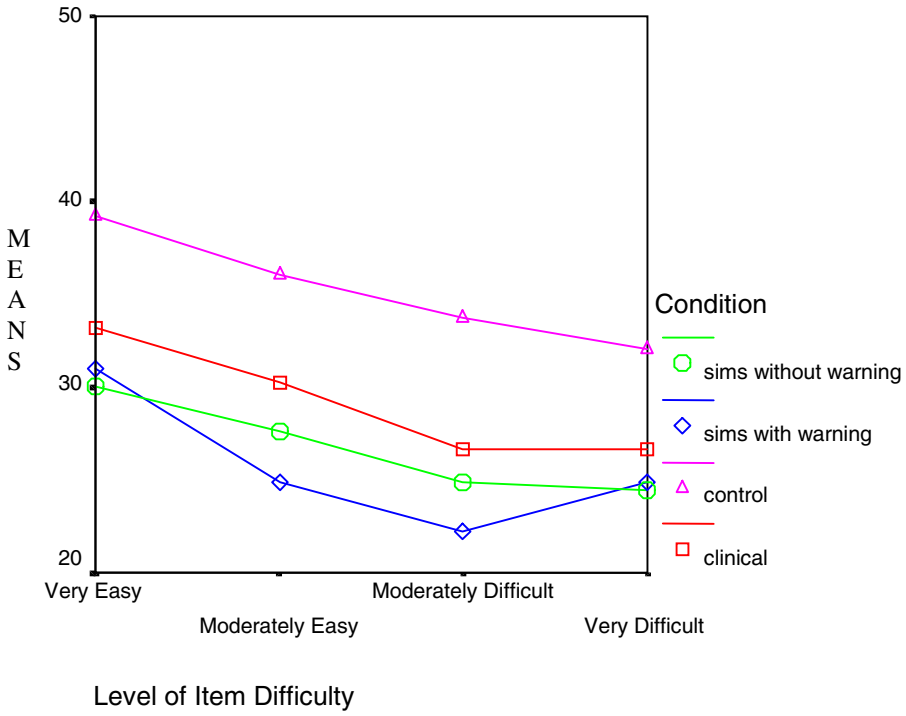


Note. Item difficulty increases from left to right. Total possible score is 104 for each level of item difficulty (i.e., 26 items at each level; each is worth a maximum of 4 points).

Figure 2 demonstrates that, in Section 1, CL and NC performance yielded qualitatively different curves than CS and NCS, particularly between Moderately Difficult and Very Difficult items. Whether these differences in curves varied by group was evaluated by calculating the mathematical difference between the Moderately Difficult and Very Difficult items and comparing the result. An ANOVA revealed that

Figure 2

Performance Curves for CS, NCS, NC, and CL on TOCA Section 1



Note. Item difficulty increases from left to right. Total possible score is 40 for each level of item difficulty (i.e., 10 items at each level; each is worth a maximum of 4 points).

groups were not significantly different, $F(3, 114) = 1.69$, $p = .17$. Similarly, the difference between Moderately Easy and Very Difficult Items was assessed. Though differences were not significant, there was a trend for the difference scores for CS to be larger and in the opposite direction from CL difference scores, $p < .08$.

Performance curves also differed in Section 3.[9] The four levels of difficulty depicted in Figure 3 demonstrate the expected negatively-accelerating curve for CL and

[9] Items from Section 2 were not used in PC scales because there were too few easy items.

NC as well as atypical curves for CS and NCS. Within-group comparisons revealed significantly different performance within each group across each level of item difficulty, $F(3, 111) = 26.47$, $p < .01$. Likewise, between-group comparisons revealed significantly different scores across item difficulty, $F(3, 111) = 12.28$, $p < .01$. Tukey's HSD demonstrated that CS and NCS were not significantly different at any level of difficulty. Table 20 displays the mean scores of each group and Cohen's d for SIMs versus CL.

Figure 3 reveals that in Section 3, CL and NC performance curves were most different between Moderately Easy and Moderately Difficult items. The difference between these particular items was negative in simulators, reflecting their positively accelerating curves. An ANOVA performed on the difference between these items revealed significant differences between groups, $F(3, 114) = 4.88$, $p = .003$. Tukey HSD located the difference to be between CS and CL ($p = .02$). CS and NCS did not demonstrate differences in likelihood of producing a positive curve between these levels, $X^2(2, n = 51) = .002$, $p > .05$.

Gudjonsson and Shackleton (1986) also calculated a metric that reflects overall rate of decay of performance curves. The formula, $(2A + B) – (D + 2E)$, where A = very easy items, B = moderately easy items, D = moderately difficult items, and E = very difficult items, computes the linear trend of rate of decay across levels of item difficulty. The formula was applied to Sections 1 and 3 and to both Sections Combined. ANOVAs and Tukey's HSD revealed significant differences on the Sections Combined scale only,

Table 20

Group Mean and Effect Size (Cohen's d) Comparisons for each Level of Item Difficulty (Performance Curve) on Section 3

	Group					
	SIM		Honest			
Level of Difficulty	CS	NCS	CL	NC	F	d
Very Easy	17.18_a (5.58)	17.37_a (6.46)	23.00_b (3.15)	23.82_b (0.85)	14.32[d]	1.28
Moderately Easy	16.45_a (5.98)	16.14_a (6.00)	21.48_b (4.05)	23.09_b (2.37)	14.22[d]	1.04
Moderately Difficult	18.09_{ab} (4.99)	17.03_a (5.31)	20.33_b (3.61)	14.55_b (3.02)	6.12[d]	0.64
Very Difficult	14.55_a (5.23)	14.97_a (5.47)	18.10_{ab} (6.42)	19.18_b (4.26)	4.23[d]	0.57

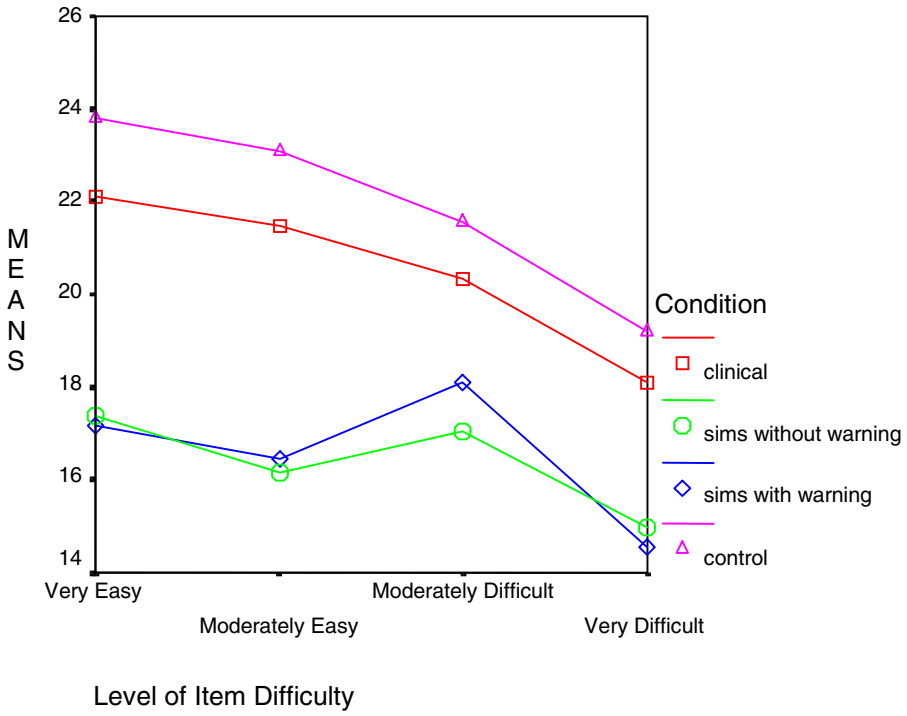
Notes. Means with different subscripts are significantly different by Tukey comparison, $p < .05$. For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls.
Given the lack of difference between CS and NCS scores, the effect size measures the magnitude of difference between mean CL score and mean SIM score. $D = M_1 - M_2$ / pooled SD.
[d] For F ratios, $p < .01$

$F(3, 111) = 7.89$, $p < .01$. While CS and NCS did not differ from each other, they both differed from CL and NC (see Table 21).

Symptom Validity Testing (SVT). The SVT strategy requires that scores be compared to those expected by chance alone. Based on a 4-alternative format, three levels of chance were used: .08 (the chance likelihood of getting both responses of a item

Figure 3

Performance Curves for CS, NCS, NC, and CL on Section 3 of the TOCA



Note. Item difficulty increases from left to right. Total possible score is 24 for each level of item difficulty (i.e., 6 items at each level; each is worth a maximum of 4 points).

Table 21

Mean Rate of Decay and Effect Size (Cohen's d) Comparisons on Sections 1, 3 and Both Sections Combined

Rate of Decay	Group: SIM CS	SIM NCS	Honest CL	Honest NC	F	d
Section 1	15.00 (16.56)	14.00 (18.27)	16.90 (17.63)	16.73 (18.96)	0.19[a]	.14
Section 3	3.64 (9.51)	3.93 (11.57)	9.14 (11.39)	10.82 (9.98)	2.93[b]	.51
Sections Combined	77.55_a (35.97)	85.31_a (53.48)	114.48_b (28.27)	120.00_b (33.30)	7.89	0.92

Notes. Cohen's d was calculated using CL versus SIM combined. For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls.

[a] p = .91

[b] p = .55

incorrect or correct), .25 (the chance likelihood of getting the first response of an item correct), and .41 (the chance likelihood of getting either the first or second response of an item correct). The next section begins with the 25%-level comparisons (for each of the three parts of the TOCA), followed by 41%, and 8%. Table 22 displays the number of participants who scored at or below each level of chance on each section of the TOCA.[10]

- SVT-First. This scale reflects the fact that in a 2-part question with 4 answer choices, there is a 25% likelihood of answering the first response correctly by chance alone. There are 60 items in Section 1, with the first response of each item worth 1 point. Thus, a score of 15 is equal to 25% of the total possible correct and is equivalent to chance performance. One CL and two SIM earned scores this low or lower but neither was significantly below chance. The fact that one non-simulating participant scored at chance raises two possibilities. First, the selection of some items may not have been equi-probable. In other words, certain items may have had a greater likelihood of being selected irrespective of being correct or incorrect. Second, some items were so difficult result in chance levels of performance, particularly in cognitively impaired participants.

SVT was applied to Sections 2 and 3 also. There are 22 first responses each worth 1 point in Section 2. A score of 5.5 is equal to 25% of 22 and is equivalent to chance

[10] The formula for significantly below chance performance is: Number of errors (E) minus the chance probability of being incorrect (Q; .92, .75, .59) x number of items (N) divided by the square root of the chance probability of being correct (P; .08, .25, .41) x chance probability of being incorrect (Q) x number of items (N). $E - QN / \sqrt{PQN}$. The yield is compared to z-scores to determine the distance beyond the normal curve (from Binder, 1992).

performance. The data in Table 22 suggest that Section 2 was so difficult that some non-clinical participants scored at chance levels when performing their best. In Section 3, there are 30 items and a score of 7.5 is 25% of the total possible. Six participants (all SIM) scored this low or lower. No respondents scored significantly below the 25% chance level in Sections 1, 2 or 3.

- SVT-Either. The second SVT comparison involved the chance likelihood of getting either response correct. There is a 41% likelihood of responding correctly to either response in a two-part question (with four response

Table 22

SVT: Number (and Percentages) of SIM, CL, and NC Who Performed at or Below Three Levels of Chance Performance on Each Section of the TOCA

	Group								
	SIM			CL			NC		
	n = 51			n = 42			n = 22		
	Level of Chance			Level of Chance			Level of Chance		
	.08	.25	.41	.08	.25	.41	.08	.25	.41
TOCA Section									
Section 1	0	2 (3.9)	4 (7.8)	0	1 (2.4)	4 (9.5)	0	0	0
Section 2	0	11 (21.6)	32 (62.7)	0	12 (28.6)	31 (73.8)	0	4 (18.2)	12 (54.5)
Section 3	0	6 (11.8)	6 (11.8)	0	0	0	0	0	0

Note. No participant scored significantly below chance on these SVT scales, ps < .05.

alternatives) by chance alone.[11] On Section 1, nine participants scored less than or equal to 98 (41% of 240 = 98);[12] five were CL and four were SIM. On Section 2, 76 participants (82.6%) scored at a level equivalent to chance (41% of 88 = 36); 32 of the 51 SIM (62.7%) scored at this level. The results from Section 2 again suggest that it is extremely difficult for most participants; 73.8% of CL and 54.5% of NC performed at or below chance on this section. On Section 3, six participants (all SIM) scored at or below chance (chance performance = 41% of 120, or 49), but none scored significantly below chance.

- SVT-Both. With 4 alternatives, there is an 8% likelihood of answering a two-part question correctly by chance alone. Thus, a mean score of 36 (8% of a possible total score of 448) was used. No individuals performed this poorly.

It is important to note that the scores that fall at or below chance performance do not demarcate significantly below chance performance. Thus, there is no reason to suspect that CL responded incorrectly on purpose. CL scores falling at the chance level likely reflect the individuals' neurocognitive limitations and the difficulty of the items.

- Truncated SVT. A fourth SVT scale was developed in which only items of medium difficulty were included. This truncated version drew from those items that were answered incorrectly by CL 25 – 50% of the time. As a result,

[11] The chance likelihood of getting either response correct is 41%, rather than 50%, because the likelihood of responding correctly to the second response is contingent upon the answer to the first response (i.e., it is a conditional probability).
[12] Section 1 contains 60 items; each is worth 4 points, for a possible total score of 240.

two groups of items were excluded: Items that most CL found too easy to miss and those too difficult not to miss. Two SIM scored significantly below the 41% chance probability level. No participants scored below the 25% level.

Research Question 2

The second research question investigated the effectiveness of each detection strategy at correctly classifying groups. Two methods were used to differentiate the participants' scores into appropriate groups: (a) cutting scores, and (b) discriminant function analysis (DFA). Each of the following strategies will be addressed individually: Floor Effect, MoE, RT, and Rate of Decay.[13] All comparisons of strategy utility involve SIM versus CL. Although four groups were used in the study, the most important comparisons are between CL and (a) NCS and (b) CS. It is not particularly useful to demonstrate that NC differs from CS and NCS because these differences do not take into account the performances of head-injured patients. Because performance did not differ significantly between NCS and CS on any scales, they were combined into one simulation group (SIM). Thus, the critical comparison was CL versus SIM. The following cutting scores and DFAs were derived for these two groups.[14]

Cutting scores were derived by performing cross-tabulations on scores from each scale by group (i.e., SIM and CL). Estimates of sensitivity, specificity, positive predictive power (PPP), negative predictive power (NPP), and hit rate were used to evaluate classificatory accuracy. As mentioned previously, cutting scores are not provided here

[13] The SVT strategy was not a metric variable and was not suitable for these analyses.
[14] For NC, classification rates ranged from 86.0% - 100%.

but can be obtained from the author by qualified researchers. The following synopsis of scores highlights meaningful differences between groups revealed by each strategy. Tables 23 – 32 provide utility estimates and results from DFAs conducted on individual scales.

NPP is generally considered the optimal score to reference when attempting to rule out conditions other than feigning. The higher the NPP, the higher the level of confidence that a participant will not be falsely labeled a patient. In the following analyses, two estimates of NPP are provided: One with NPP optimized (to minimize false positives) and one based on optimal cutting scores overall. The second set of utility estimates was computed for each strategy without particular regard to NPP but with an effort to produce the best overall classification rates.

The following sections provide the results from cross-tabulations and DFAs performed on each strategy. The majority of the data are presented in tabular form, while an overview of the results is provided in text. For each strategy, the results from cross-tabulations (i.e., cutting scores) will be provided first, followed by the results from DFA. Each strategy is reviewed individually.

Section Score. Section Scores reflect the total number of items correct on each section of the TOCA; thus, four scales were calculated (Sections 1, 2, 3, and Total Score). The extent to which cutting scores accurately discriminated groups when using Section Scores was examined (see Table 23). Section 3 scores below the cutting score yielded the best classification (75.0%), whereas Section 2 score yielded the poorest classification

Table 23

Utility Estimates for Classifying SIM versus CL Derived from Optimal Hit Rates Using Four Section Scores

	Utility Estimates (%)				
Condition	Sens	Spec	PPP	NPP	HR
Section 1	72.6	68.3	74.0	67.7	70.7
Section 2[a]	66.7	22.7	50.7	37.0	47.8
Section 3	80.4	68.3	75.9	74.0	75.0
Total Score	84.3	56.1	70.5	74.2	71.7

Notes. Sens = Sensitivity; Spec = Specificity; PPP = Positive Predictive Power; NNP = Negative Predictive Power; HR = Hit Rate. For Groups, CL = Clinical Group, SIM = Both Simulating Groups.

[a] In contrast to the other section scores, scores for CL in Section 2 were lower on average than those for SIM. Utility estimates were calculated from scores above the cutting score.

(47.8%). Increasing the NPP to its maximum, resulted in improved sensitivity only at the cost of very poor specificity and PPP. Hit rates were also generally lower.

The second analysis utilized Discriminant Function Analysis (DFA) to identify regression formulas that statistically maximize the variance between groups relative to the variance within groups (Hair, Anderson, Tatham, & Black, 1995). The extent to which each Section Score and Total Score correctly differentiated groups was analyzed. As can be seen in Table 24, Section 3 yielded the best classification rate (72.0%), while Section 2 yielded the poorest rate (61.3%).

Floor Effect. The goal of the Floor Effect strategy is to differentiate feigners from brain-injured participants by identifying respondents who answer very easy items incorrectly. Three Floor Effect scales were evaluated. In terms of cutting scores, the 22-response version of the Floor Effect (i.e., FE-95%-Responses) produced the best overall classification rate, while the 7-item version (FE-95%-Items) yielded the best PPP

Table 24

Classifications (and Percentages) of SIM versus CL Based on Direct Discriminant Function Analysis on Each Section Score

	Actual Group Membership [a]		
Predicted	CL	SIM	Classification Rate (%)
Section 1			
CL	23 (56.6)	12 (23.5)	67.4
SIM	18 (43.4)	39 (76.5)	
	n = 41[b]	n = 51	
Section 2			
CL	23 (54.8)	17 (33.3)	61.3
SIM	19 (45.2)	34 (66.7)	
	n = 42	n = 51	
Section 3			
CL	32 (76.2)	16 (31.4)	72.0
SIM	10 (23.8)	35 (68.6)	
	n = 42	n = 51	
Total Score			
CL	23 (54.8)	11 (21.6)	65.7
SIM	18 (45.2)	40 (78.4)	
	n = 41[b]	n = 51	

Note. For Groups, CL = Clinical and SIM = Simulators Combined (CS and NCS).
[a] Percentages are shown in parentheses.
[b] One participant did not complete Section 1 but did complete the other sections.

(see Table 25). DFAs on each of the Floor Effect scales revealed that the FE-95%-Item scale provided the best overall classification rate, while the FE-95%-22-Rresponse and FE-90%-Item scales were identical in their classificatory accuracy (see Table 26). The FE-95%-Item scale yielded a particularly favorable proportion of true negatives (85.4%).

Table 25

Utility Estimates for Classifying SIM versus CL Derived from Hit Rates Using Three Different Floor Effect Scales

	Utility Estimates (%)				
Floor Effect Scale	Sens	Spec	PPP	NPP	HR
FE-95% (7 Items)	72.6	85.4	86.1	71.4	78.3
FE-90% (13 Items)	80.4	78.0	82.0	76.2	79.4
FE-95% (22 Responses)	80.4	80.5	83.7	76.7	80.4

Note. Sens = Sensitivity; Spec = Specificity; PPP = Positive Predictive Power; NNP = Negative Predictive Power; HR = Hit Rate. NPP could not be increased on these scales.

Table 26

Classifications (and Percentages) of SIM versus CL Based on Discriminant Function Analysis on Three Floor Effect Strategies

	Actual Group Membership [a]		
Predicted	CL	SIM	Classification Rate
	n = 41	n = 51	
FE-95% (7-item)			
CL	35 (85.4)	14 (28.3)	78.3
SIM	6 (14.6)	37 (71.7)	
FE-90% (13-item)			
CL	34 (82.9)	19 (37.3)	71.7
SIM	7 (17.1)	32 (62.7)	
FE-95% (22-response)			
CL	34 (82.9)	19 (37.3)	71.7
SIM	7 (17.1)	32 (62.7)	

Note. For Groups, CL = Clinical and SIM = Both Groups of Simulators.
[a] Percentages are shown in parentheses.

Attempts to increase NPP were unsuccessful; thus the other utility estimates remained unchanged.

Magnitude of Error. The extent to which the MoE strategy effectively discriminated CL from SIM was evaluated via cutting scores and DFA. For the truncated MoE scale (i.e., 25 – 50% incorrect items), scores below the cutting score correctly classified 68.5% of the participants (see Table 27).

Higher scores were indicative of feigning on the 7.1%-MoE scales. Higher scores reflect the increased tendency to miss items that CL rarely miss. On Section 1, scores above the cutting score correctly classified 76.1%, while on Section 3 scores higher than the cutting score correctly classified 93.5%.

Combining the two 7.1%-MoE scales correctly classified 82.0%. The 7.1%-MoE3 scale yielded the best results overall. It also produced excellent sensitivity and PPP without concurrent reductions in specificity and NPP. Increasing the NPP improved sensitivity on the Section 1 and Combined Sections scales but also reduced specificity substantially.

DFAs were conducted on each approach to the MoE strategy on Section 1. The Truncated scale resulted in a classification rate of 68.5%. The DFA performed on 7.1%-MoE3 provided the best discrimination, correctly classifying 84.8% of the participants. The true negative rate was particularly good (92.7%).

Response Time (RT). The RT strategy examined the extent to which SIM delayed their responses in order to appear head-injured. RT was the average number of minutes to

Table 27

Utility Estimates for Classifying SIM versus CL Derived from: (a) Optimized Hit Rates and (b) Optimized NPP Using Three Magnitude of Error (MoE) Scales

		Optimized Hit Rates (%)			
MoE Scale	Sens	Spec	PPP	NPP	HR
7.1%-MoE1	70.6	78.0	80.0	68.1	76.1
7.1%-MoE3	94.1	92.7	94.1	92.7	93.5
7.1%-MoE1+3	88.2	73.2	80.4	83.3	82.0
Truncated MoE	76.5	58.5	69.6	66.7	68.5
		Optimized NPP (%)			
7.1%-MoE1	92.1	44.0	66.2	81.0	68.5
7.1%-MoE3	94.1	92.7	94.1	92.7	93.5
7.1%-MoE1+3	96.1	44.0	86.0	90.0	73.0
Truncated MoE	96.1	12.2	67.7	80.0	69.6

Notes. Sens = Sensitivity; Spec = Specificity; PPP = Positive Predictive Power; NNP = Negative Predictive Power; HR = Hit Rate. Section 2 scores were not used.
For Groups, CL = Clinical and SIM = Both Groups of Simulators.

complete a section. On Section 2 the computer enforced a time limit (40 seconds per item). Table 29 provides utility estimates optimizing both NPP and overall hit rate for the RT strategy.

Table 28

Classification (and Percentages) of SIM versus CL Derived from Discriminant Function Analysis Using Four MoE Scales

	Actual Group Membership [a]		
Predicted	CL	SIM	Classification Rate (%)
7.1%-MoE1			
CL	31 (75.6)	16 (27.5)	73.9
SIM	10 (24.4)	37 (72.5)	
	n = 41	n = 51	
7.1%-MoE3			
CL	39 (92.7)	11 (21.6)	84.8
SIM	3 (7.3)	40 (78.4)	
	n = 42	n = 51	
7.1%-MoE1+3			
CL	33 (80.5)	14 (23.5)	78.3
SIM	8 (19.5)	39 (76.5)	
	n = 41	n = 51	
Truncated MoE			
CL	24 (58.5)	12 (23.5)	68.5
SIM	18 (41.5)	39 (76.5)	
	n = 42	n = 51	

Notes. For Groups, CL = Clinical and SIM = Both Groups of Simulators.
The "7.1%" scales were derived by identifying incorrect responses chosen by fewer than 5 CL in Sections 1 and 3. No items from Section 2 met this criteria. The truncated MoE scale was derived by (a) identifying those items that 25 – 50% of the CL answered incorrectly, and (b) identifying incorrect responses to the above items chosen by < 10% of CL.
[a] Percentages are shown in parentheses.

Table 29

Utility Estimates for Classifying SIM versus CL Derived from Hit Rates Using Four Response Time (RT) Scales

		Optimized Hit Rates (%)			
Condition	Sens	Spec	PPP	NPP	HR
RT1	86.3	80.4	86.3	82.9	83.7
RT2	82.4	61.9	72.4	74.3	73.9
RT3	66.7	64.3	69.4	67.5	66.3
RT Total	88.2	78.1	83.3	84.2	84.8
		Optimized NPP (%)			
RT1	91.0	54.8	57.6	93.2	78.5
RT2	45.2	33.5	47.6	81.2	46.5
RT3	87.1	36.3	65.2	79.5	75.9
RT Total	98.3	25.3	68.7	93.2	80.1

Note. Sens = Sensitivity; Spec = Specificity; PPP = Positive Predictive Power; NNP = Negative Predictive Power; HR = Hit Rate.

Scores above the cutting score on time to complete the TOCA (Total RT) classified 84.8% of the participants correctly. Scores above the cutting score on Section 1 were also effective discriminators, correctly classifying 83.7%, while RTs from Sections 2 and 3 were less efficient. As can be seen in Table 30, the DFAs performed on Section 1 RT and on Total RT produced the best hit rates.

Rate of Decay. The Rate of Decay is a quantitative measure of the performance curve for each participant (i.e., the overall linear trend of decay across item difficulty). A relatively higher rate is expected in honest responders and a relatively lower rate is expected in feigners. Such a pattern of findings is expected because feigners are thought to be more likely to produce flatter curves, which result in lower slopes (i.e., rates of

Table 30

Classifications (and Percentages) of SIM versus CL Derived from Discriminant Function Analysis on Four RT Scores

	Actual Group Membership [a]		
Predicted	CL	SIM	Classification Rate (%)
RT1			
CL	26 (65.0)	3 (5.9)	81.5
SIMS	14 (35.0)	48 (94.1)	
	n = 41	n = 51	
RT2			
CL	27 (64.3)	10 (19.6)	73.1
SIM	15 (35.7)	41 (80.4)	
	n = 42	n = 51	
RT3			
CL	13 (31.0)	7 (13.7)	61.3
SIM	29 (69.0)	44 (86.3)	
	n = 42	n = 51	
RT Total			
CL	26 (65.0)	1 (2.0)	83.7
SIM	15 (35.0)	50 (98.0)	
	n = 41	n = 51	

Note. For Groups, CL = Clinical and SIM = Simulators Combined (CS and NCS).
[a] Percentages are shown in parentheses.

decay). Section 2 was not included when the sections were combined due to the lack of significant findings when it was included.

Scores falling below the cutting score for Rate of Decay on Sections 1 and 3 of the TOCA correctly classified 71.7% of participants (see Table 31). Increasing the NPP to 68.5% produced an increase in sensitivity and a decrease in specificity, PPP, and hit rate. DFA on the overall rate of decay on the TOCA classified 66.3% of the participants correctly (see Table 32).

Table 31

Utility Estimates for Classifying SIM versus CL Derived from Hit Rates Using Rate of Decay Strategy on Sections 1 and 3 Combined

	Utility Estimates (%)				
Rate of Decay	Sens	Spec	PPP	NPP	HR
Sections 1 & 3 Combined	64.7	73.2	75.0	62.5	71.7

Note. Sens = Sensitivity; Spec = Specificity; PPP = Positive Predictive Power; NNP = Negative Predictive Power; HR = Hit Rate. Section 2 was not included in the analysis due to excessive variability in responses (levels of difficulty could not be established). Rates of Decay for Sections 1 and 3 used individually are not included due to the lack of significant differences among groups on those variables. NPP could not increased appreciably on this scale.

Table 32

Utility Estimates for Classifying SIM versus CL Derived from Discriminant Function Analysis on Rate of Decay Scale

	Actual Group Membership [a]		
Predicted	CL	SIM	Classification Rate (%)
CL	25 (61.0)	15 (29.4)	66.9
SIM	16 (39.0)	36 (70.6)	
	n = 41	n = 51	

Note. For Groups, CL = Clinical and SIM = Simulators Combined (CS and NCS). Section 2 was not included in the analysis due to excessive variability in responses (levels of difficulty could not be established). Rates of Decay for Sections 1 and 3 used individually are not included due to the lack of significant differences among groups on those variables.

[a] Percentages are shown in parentheses.

Other Differences in PC. The extent to which differences at particular segments of the curves correctly classified simulators was also examined. The most effective comparison was between Moderately Easy and Moderately Difficult items on Section 3.

Scores below the cutting score while optimizing overall Hit Rate, correctly classified 66.7%, with a sensitivity of 78.4% and specificity of 52.4%. Attempts to optimize NPP failed to increase NPP. A DFA also correctly classified 66.7%.

Research Question 3

The third research question investigated the utility of combining strategies when differentiating SIM from CL. In the previous section (Research Question 2), classification rates of each scale were identified. The following section explores whether combining strategies leads to incremental validity. The effect of combining strategies on classificatory accuracy was measured by DFA.

First, based on DFA results, the most effective scales from each strategy were selected; namely: RT Total, FE-95%-Item, 7.1%-MoE3, and Overall Rate of Decay. Second, the variables were entered into a Stepwise DFA. The RT Total and 7.1%-MoE3 scales were entered and resulted in a classification rate of 95.7% (see Tables 33 & 34). This classification rate represents a slight increase from that of the most effective strategy used alone (7.1%-MoE3; 84.8% when using cutting scores and 93.5% when using DFA). Strategies for which metric variables could not be constructed (i.e., SVT and PC strategies) were not used in these analyses.

Thus, combining strategies resulted in an improvement in classification rate, a reduction in the number of false negatives, and a less dramatic but important reduction in number of false positives. As will be discussed in the following chapter, the finding that combining strategies improves classificatory accuracy is consistent with the work of Rogers (1997) and Frederick and Foster (1991).

Table 33

Group Classifications and Percentages from Stepwise DFA on the Four Most Effective Scales: RT Total, 7.1%-MoE3, FE-95%, and Rate of Decay[a]

	Actual Group Membership [b]		
Predicted	CL	SIM	Classification Rate (%)
CL	39 (95.1)	2 (3.9)	95.7
SIM	2 (4.9)	49 (96.1)	
	n = 41	n = 51	

Note. For Groups, CL = Clinical and SIM = Simulators Combined (CS and NCS). FE-95% and Rate of Decay variables were not entered.
[a] Only the 7.1% MoE3 and RT Total variables were entered into the function based on Wilk's Lambda > 1.00.
[b] Percentages are shown in parentheses.

Table 34

Standardized Canonical Discriminant Function Coefficients and Structural Correlations for Stepwise DFA on the Four Most Effective Detection Strategies

Strategy	Canonical Coefficient	Structural Correlation [b]
7.1%-MoE3	.77	.75
RT Total	-.66	.63
FE-95%-Items[a]	N/A	-.24
Rate of Decay[a]	N/A	-.24

[a] This variable not used in the analysis.
[b] Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions.

To evaluate the degree to which classification may drop when all variables are entered into the function, a Forced-Entry DFA was performed. The results show that 94.6% of CL and SIM were correctly classified (see Table 35). Table 36 displays the

Table 35

The Effect of Combining the Four Most Effective Scales: Group Classifications and Percentages from Forced-Entry DFA on RT Total, 7.1%-MoE3, FE-95%, and Rate of Decay

	Actual Group Membership [a]		
Predicted	CL	SIM	Classification Rate (%)
CL	39 (95.1)	3 (5.9)	94.6
SIM	2 (4.9)	48 (94.1)	
	n = 41	n = 51	

Note. For Groups, CL = Clinical and SIM = Simulators Combined (CS and NCS).
[a] Percentages are shown in parentheses.

Table 36

Standardized Canonical Discriminant Function Coefficients and Structural Correlations for Forced-Entry Method DFA on the Four Most Effective Detection Strategies

Strategy	Canonical Coefficient	Structural Correlation [a]
7.1%-MoE3	-.69	-.74
RT Total	.63	.63
FE-95%-Items	.11	.41
Rate of Decay	.13	.42

[a] Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions.

relative contributions of each scale to classification. The 7.1%-MoE3 scale accounted for the most classificatory power of the function. Total RT accounted for the next most, while the 7-Item Floor Effect and Rate of Decay strategies contributed relatively less to the function.

Analyses on Less Impaired CL. Additional analyses were conducted on a subset of less impaired CL. The rationale for these analyses was that the majority of cases in

neuropsychology that involve a determination of malingering requires differentiation of feigning from mild injuries. A common referral question is to determine neurocognitive status following a minor car accident in which a mild TBI may or may not have been sustained. Most individuals in these situations are not hospitalized for acute injuries, otherwise their injuries would not be considered "mild." In order to evaluate the utility of the detection strategies in a mildly impaired sample, severely impaired patients in the clinical group were removed and the remaining 11 patients with mTBI were compared to SIM. The overall classification rate fell slightly, from 94.6% to 90.3%. The overall reduction in classification was due to a substantial increase in the number of false positives (5 of the 11 mTBI were misclassified as SIM). However, the true positive rate (the number of SIM correctly classified as such) increased to 98.0%. Given the very small number of mTBI patients in the analysis, the DFA was likely unstable; these data should be interpreted cautiously.

The Effect of Base Rate on Classification. Sweet et al. (2000) and Rosenfeld, Sands, and Van Gorp (2000) have argued that studies of feigning should not only include utility estimates based on group composition and bases rates employed in the study, but also utilize more real-world estimates of prevalence. Debate still exists about the prevalence of malingering and several variables appear to be associated with a given local base rate (e.g., medicolegal context, setting, diagnosis). The sample of simulators used in the current study (i.e., 48.4% of all participants were SIM) is disproportionately large as compared to most estimates. As a result, the classification rates derived from the study may not be as accurate as those for which a lower prevalence is assumed. In

particular, the results may provide overestimates of PPP. A DFA was performed on a sample of SIM chosen at random to reflect a base rate of 15.0% (14 SIM), which is in line with the recommendations of Sweet et al. (2000). The four most effective strategies were again used in the DFA.

The classification rate based on a more representative base rate of malingering (15.0%) fell slightly from that of the higher base rate (Table 37). However, the number of false positives remained quite low (4.9%). The 7.1%-MoE3 again contributed the most to the function (see Table 38).

Research Question 4

The final research question addressed the effect of cautioning simulators that their performance would be monitored for attempts to feign impairment. Although CS performance did not differ statistically from that of NCS on any of the scales, two non-significant (*p*s > .05) trends were noted. First, CS scored slightly higher than NCS on the the 7.1%-MoE scales. Scoring higher on these scales is consistent with what was expected from simulators. Second, NCS performed more like NC than CS did on each level of item difficulty except Very Difficult items (see Table 17). Neither finding is consistent with the hypothesis that CS, in an effort to feign believably, would be more cautious and perform more like NC. In fact, CS scores were higher than NCS scores on most scales (12 for CS versus 9 for NCS). Such a finding suggests that most CS scores were actually more indicative of feigned performance than NCS scores. It may be that the caution led CS to believe that their chances of going undetected were poor. As a result, they may have put forth less effort to feign believably.

Table 37

Group Classifications and Percentages from Discriminant Function Analysis on SIM versus CL with the Base Rate of SIM Set at 15.0%

	Actual Group Membership [a]		
Predicted	CL	SIM	Classification Rate (%)
CL	39 (95.1)	3 (21.4)	90.9
SIM	2 (4.9)	11 (78.6)	
	n = 41	n = 14 [b]	

Note. For Groups, CL = Clinical and SIM = Simulators Combined (CS and NCS).
[a] Percentages are shown in parentheses.
[b] Fourteen SIM (15.0%) were chosen at random to reflect a more representative base rate of malingering (Sweet, 2000).

Table 38

Standardized Canonical Discriminant Function Coefficients and Structural Correlations for Forced-Entry Method DFA on the Four Most Effective Detection Scales with Base Rate Equal to 15.0%

Strategy	Canonical Coefficient	Structural Correlation [a]
7.1%-MoE3	.91	.84
RT Total	-.30	-.25
FE-95%-Items	-.14	-.24
Rate of Decay	-.39	-.36

[a] Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions.

Summary of Results

The results for the first research question suggested that NC do indeed differ from CL and SIM on the TOCA. Each scale revealed a predicted pattern, with elevated scores for feigners. The second research question produced results that support strategic detection of feigning. Utility estimates ranged from modest to excellent, with the MoE strategy showing particular success. The subsequent section demonstrated that combining scales provides the best classification rates, with 7.1%-MoE3 and RT Total providing the most discriminatory power. Finally, the fourth research question suggested that cautioning simulators that they will be monitored for feigning on the TOCA had little effect on their performance. CS may have thought it unlikely that they could go undetected due to the presence of the detection strategies. Consequently, they may have been poorly motivated to feign credibly. Similarities and contrasts of these findings to previous research will be discussed in the following chapter.

CHAPTER IV

DISCUSSION

A patient's best effort during a neuropsychological evaluation is critical to the validity and interpretation of the results. Within the past twenty years, the vulnerability of neuropsychological tests to the effects of incomplete effort, poor motivation, feigning, or malingering has been well established. During the same time, the rewards available to plaintiffs who demonstrate neurocognitive compromise have increased steadily, making exaggeration of symptoms (or malingering) more appealing to plaintiffs. Numerous studies have explored the utility both of new measures designed specifically to detect feigning and of feigning scales developed from existing measures of neurocognitive function. However, methodological, statistical, and practical limitations have prevented the development of a single test or index that consistently detects feigners without sometimes incorrectly identifying patients as feigners (i.e., yielding false positives). False positives represent a major limitation in assessment of neurocognitive feigning and while certain tests have been shown to differentiate feigners from patients, they do not by themselves stipulate the cause of feigning. Classification of malingering ultimately requires clinical judgment. Given the potentially harmful consequences of falsely labeling a patient as a malingerer, research aimed at developing new strategies, refining

existing strategies, and improving classification rates is essential to the practice of clinical neuropsychology.

The present study addresses two major goals: (a) an examination of the utility of five detection strategies, and (b) a preliminary validation of a new measure of neurocognitive feigning, the TOCA (Rogers, 1996). The TOCA is comprised of three sections, each of which appears to measure a different neurocognitive ability. It also utilizes several feigning detection scales, which were developed and evaluated in the current study. Pilot data suggest that the TOCA may also be a measure of cognitive ability. The dual emphasis on cognitive ability and feigning places the TOCA in a position to provide useful information regardless of feigning status. The current project focused on the utility of the TOCA as a measure of feigning.

This Discussion begins with a summary of the primary findings of the study and how they are categorized by Rogers' (1997) model of classificatory certitude. The hit rates and DFA classification rates are provided for each scale to facilitate direct comparisons. Following the overview, the Discussion is organized into two major sections.

The first major section of this chapter addresses theoretical and methodological issues: (a) the conceptual framework of feigning detection and the characteristics of effective and ineffective strategies, and (b) threats to external validity inherent to simulation design. These themes constrain the interpretability of data obtained from current research methods in several ways. In addition, the issues of instruction sets, warnings, coaching, incentives, and posttest manipulation checks are discussed.

The second major section reviews the practical implications of past research and the current findings. Problems inherent to classification of feigning and their impact on clinical practice are addressed. The utility of potential diagnostic criteria (Slick et al., 1999) and threshold models (Pankratz & Binder, 1997) for malingering are briefly discussed.

The concluding section contains a summary of the study's strengths, weaknesses, and proposals for future research. For example, this study addresses several important methodological considerations, including instructional sets that emphasized credibility and symptom specificity. A manipulation check of participants' comprehension of and compliance with the instructions was also performed. Possibly the most important consideration was the evaluation of multiple detection strategies, both in isolation and in combination. Limitations of the study include the reliance on simulating college students and the lack of a formal estimate of intellectual level. Recommendations for future research include further exploration of the utility of combining strategies, particularly MoE, FE, and RT.

Primary Findings of the Study

The main objective of this study was to evaluate the effectiveness of five feigning detection strategies, used in isolation and in combination, at classifying simulators. In the initial stages of scale development, it is important to assess the strategies from multiple perspectives. Thus, cutting scores were operationalized both on rational and empirical grounds. Also, the effectiveness of each scale was assessed by using two methods: First, cutting scores on each scale that optimally classified simulators and patients were

established. This procedure allowed for evaluation of utility estimates (i.e., sensitivity, specificity, positive predictive power, negative predictive power, and hit rate). Second, Discriminant Function Analyses (DFA) provided classification rates of simulators and patients based on standardized canonical coefficients. DFA differentiates groups according to the ratio of between-group to within-group variance. After testing individual scales, the effect of combining scales was evaluated by DFA.

The results indicated that Floor Effect, Magnitude of Error, Performance Curve, Response Time, and Symptom Validity Testing strategies, as operationalized for this study, ranged in clinical utility from Speculative to Definite (Rogers, 1997). The MoE strategy was most effective, followed by RT, Floor Effect, PC, and SVT.

Table 39 displays the correlations between the four most effective scales. As can be seen in the table, the RT Total scale was the only scale not at least moderately correlated with another scale. The highest correlation was observed between the Rate of Decay and FE-95% scales, while the lowest was between RT Total and Rate of Decay. The higher correlations bring into question whether the scales should be construed as individual scales. It may be that they reflect variations of a larger underlying construct. In contrast, RT appears to represent another construct.

The utility estimates often depended on the section in which the strategy was used or on the specific operationalization. For example, the MoE strategy was particularly effective on Section 3 but less so on Section 1. The idea that the content of the task (e.g., verbal or nonverbal content) in which the scale is used is an important issue that will be discussed in subsequent sections. As summarized in Table 40, the 7.1%-MoE

Table 39

Correlational Matrix of Four Major Detection Scales.

Scale	7.1%-MoE3	RT Total	FE-95%-Items
Rate of Decay	-.48[a]	.23[b]	.65[a]
7.1%-MoE3		-.29[b]	-.52[a]
RT Total			.24[b]

[a] Correlation is significant at the .01 level (2-tailed).
[b] Correlation is significant at the .05 level (2-tailed).

operationalization worked better than the truncated MoE scale. Overall, the strategies were particularly effective when used in combination, achieving a classification rate of over 95%. Table 40 summarizes classification rates and level of certitude for each scale.

The Conceptual Framework of Detection Strategies

Early research in malingering paid little heed to theoretical or conceptual bases for feigning detection. Rather, studies relied on empirical and intuitively appealing findings, irrespective of the possible underlying constructs. These studies led to a proliferation of poorly focused and heterogeneous findings. Rogers, Harrell, and Liff (1993) were the first to attempt a systematic, conceptually driven review of feigning detection. They discussed common measures of feigning in terms of six detection strategies: Magnitude of Error (MoE), Floor Effect, Performance Curve (PC), Symptom Validity Testing (SVT), Atypical Performance, and Psychological Sequelae. The present study examined the utility of the first four of these strategies and RT. The following

Table 40

Classification Rates by Decision Rule and Discriminant Function, and Relative Success of Each Scale according to Rogers' (1997) Qualitative Classifications

	% Classified Correctly by		Level of Certitude
	Decision Rule	DFA	
4 Best Scales			
7.1%-MoE3	93.5	84.3	Probable/Definite
RT Total	85.0	82.0	Probable
FE-95%-Item	78.3	78.3	Probable
Rate of Decay	71.7	66.9	Tentative
Other Scales			
7.1%-MoE1	76.1	73.9	Probable
7.1%MoE1+3	84.8	79.1	Probable
Truncated MoE	68.6	68.5	Tentative
FE-95%-Response	80.4	71.7	Probable
FE-90%-Item	79.4	71.7	Probable
RT1	83.7	82.1	Probable
RT2	73.9	73.1	Tentative
RT3	66.3	61.3	Tentative
PC[a]	66.7	66.7	Tentative
SVT[b]	47.3		Speculative
2 Best Scales Combined (Stepwise DFA)		95.7	Definite
4 Best Scales Combined (Forced-Entry DFA)		94.6	Definite

Notes. The level of certitude is based on this study only.
The most effective scale from each of the 4 metric strategies (MoE, Floor Effect, Rate of Decay, and RT) was included in the DFA. Other scales may have had better hit or classification rates when used by themselves, but they provided less incremental validity than the scales chosen for inclusion in the DFA.
[a] The Performance Curve strategy does not yield utility estimates, rather it provides a visual depiction of performance that can be compared to an expected curve.
[b] SVT yielded a sensitivity of 4.0% and specificity of 100%.

section summarizes the constructs of these strategies and explores the conceptual implications of past and current findings.

Magnitude of Error

Martin, Franzen, and Orey (1998) recently demonstrated the potential of detection of feigned neurocognitive impairment by examining the degree to which a response is wrong. The current results regarding 7.1%-MoE support their findings. Unexpectedly, limiting MoE items to those answered incorrectly by 25 - 50% of CL (i.e., Truncated MoE) did not work as well. The relative failure of this version of the strategy is surprising given its intuitive appeal. First, the Truncated MoE scale utilized medium difficulty items only, thus removing items at the extremes that differentiate groups poorly. Second, it used incorrect responses chosen by less than 10% of CL. Regarding the first rule, it may be that the stipulation was too stringent. In other words, requiring that the items had to be missed by 25 - 50% of CL in order to be included may have produced a collection of slightly more difficult items than anticipated. Easier items (e.g., 15% - 35% of CL answered items incorrectly) may be more effective. However, discriminatory power will likely be lost if items are so easy that few feigners miss them. As a result, using items from a range between very easy and moderately easy appears warranted.

The 7.1%-MoE strategy worked very well, especially on Section 3. It seems likely that the relative ease of Section 3 facilitated simulators' attempts to appear impaired by allowing them to identify the correct response and choose an incorrect response. Alternatively, the findings of Boone et al. (2000) suggest that verbal, overlearned content is particularly conducive to feigning detection (see the subsequent section on overlearned

material). Specifically, SIM may have overestimated the vulnerability of overlearned material to brain injury and performed disproportionately poorly. If this is so, then simulators would be expected to be vulnerable to detection on other scales in Section 3. In testing this hypothesis, the PC strategy in Section 3 was the most effective use of PC. This trend did not hold true for RT in Section 3, which was less effective than RTs in the other sections.

Based on the percentage of items correct in each section, Section 1 was more difficult than Section 3. Therefore, it may have been more difficult for SIM to identify commonly chosen, yet incorrect responses. Moreover, Section 1 requires significant working memory skill. Such sustained attentional demands are susceptible to fluctuations in concentration, which are in turn sensitive to a variety of conditions, including brain injury and emotional distress. This vulnerability likely contributed to the variability of scores in Section 1. Similarly, the visual analysis and conceptual reasoning requirements of Section 2 are quite sensitive to neuropsychological compromise; they may have reduced discriminatory power of some of the detection scales. Section 2 was also very difficult for most participants. Items from Section 2 did not contribute to the MoE scales apparently due to their disproportionate difficulty.

The TOCA compared favorably with the only other study that employed the Magnitude of Error strategy. Martin, Franzen, and Orey (1998) reported that 86.0% of simulators and 80.0% of brain-injured patients were correctly classified by an MoE strategy applied to memory tests. The TOCA correctly classified 94.1% of simulators and 92.7% of the clinical group.

The Floor Effect

Multiple operationalizations of the Floor Effect exist in this large body of research. However, all variations capitalize on a tendency for feigners to miss items that most impaired persons answer correctly (Nies & Sweet, 1994; Rogers et al., 1993). In past research, classification rates tend to be quite good for the strategy, with the best being reported for the TOMM (Tombaugh, 1997). The TOMM utilizes extremely easy items that 99.0% of a clinical sample answered correctly. It is ostensibly a nonverbal test of recognition memory. The Floor Effect scales of TOCA, though somewhat less efficient, have advantages of being just one of several scales and covering more than one cognitive domain.

Other studies (e.g., Tenhula & Sweet, 1996; Tombaugh, 1997) have reported high classification rates when using the Floor Effect strategy. One reason for their success may be the use of long Floor Effect scales (e.g., the TOMM has 50 items). A second potential reason is the use of homogenous clinical samples that narrows the amount of variability in CL responses, which likely leads to improved discrimination from SIM. Also, these measures are positively skewed, meaning that normals do not produce normal curves. Rather, they have little problem achieving high scores. This increases the effectiveness of a strategy, such as the Floor Effect, that depends on the relative ease of the task. Several problems exist with these methods. First, homogeneous groups are not characteristic of the patients seen in clinical practice. Second, tests comprised of easy items only are more susceptible to detection by feigners, who may perform within expectations on the simple task but feign impairment on more difficult tasks.

Floor Effect scales were not developed for each section individually. Due to the a priori decision to use items that 90.0% and 95.0% of patients answered correctly, each section scale would have had insufficient items. However, given the positive results in the current study, alternate operationalizations that allow for the Floor Effect to be applied to each section appear warranted. The relative ease of Section 3 makes it particularly conducive to this strategy.

Performance Curve

Prior research has shown that atypical Performance Curves (PC; i.e., flat or positively accelerating curves) across items of increasing difficulty may indicate feigning (Frederick & Foster, 1991). "Normal" curves reflect the increases in item difficulty, and therefore accelerate negatively.

The PCs for SIM in Section 3 were clearly atypical. In the middle of the curve SIM showed an increase in performance as item difficulty increased. As would be expected, NC and CL PCs did not show such an increase. Thus, although not a metric comparison of means, the PC strategy saliently revealed a highly implausible pattern of responding. PC for Section 1 was less effective because SIM scored lower on average at each level of difficulty.

Future research implementing the PC strategy may benefit from using more levels of item difficulty. Frederick and Foster (1991) employed 10 levels to plot PCs, derived from running means. Using more levels may make the curves more sensitive to changes in performance.

The rate of decay strategy was less effective as compared to the work of Gudjonsson and Shackleton (1986). Methodological limitations of their study may explain the difference in effectiveness. The authors instructed the participants simply to fake "substantially and convincingly" below their genuine ability. A scenario was not used. No mention was made of the importance of credibility, which may have resulted in exaggerated deficits and greater differentiation from the clinical group. Alternatively, the difference in effectiveness may at least partially be due to the fact that the Raven's assesses only one cognitive domain, which is easier than the TOCA. Overall, differences in performance curves are difficult to compare given the differences in the measures themselves.

Difference scores between certain adjacent levels of difficulty (also based on the Rate of Decay strategy) revealed differences between simulators and patients. Easier items are answered correctly more often and should yield higher scores than more difficult items. Thus, when the score obtained on more difficult items is subtracted from the score on easier items, the difference should be positive. The difference scores were often negative for simulators, which is consistent with positive accelerating curves (i.e., better performance on more difficult items). The approach of examining adjacent levels of difficulty on the curve allows for the evaluation of segments of the curve rather than the curve as a whole. With more levels of item difficulty, this strategy may gain utility due to the increase in number of comparisons available.

The Rate of Decay strategy may obscure some of the differences within the curve because it is based on one score based on the entire curve. The Rate of Decay strategy is

intuitively appealing and has produced positive results when differentiating feigners from patients. Thus, it remains an attractive strategy for future studies. As will be discussed, Rate of Decay may have particular utility when used in combination with RT.

Response Time

The potential utility of Response Time in feigning research recently has been proposed but reports have been mixed as to the effectiveness of RT (Allen et al., 1997; Beetar & Williams, 1995; Binks et al., 1997; Rose et al., 1995). These differences appear to be related to content of the tasks being measured. Generally, RT on easier tasks (e.g., object recognition or dot counting) has produced better discriminatory results than RT on more difficult tasks (e.g., delayed free recall). RT on the TOCA yielded excellent hit rates, which are consistent with very recent evidence supporting the potential of RT (Boone et al., 2000; Davis et al., 1997; Rees et al., 1999). The RT scales differentiated groups well because RT was relatively low in SIM and high in CL. This result is consistent with previous findings of higher RT in CL on the PDRT (Rose et al., 1995). However, the current data are contrary to the RT data reported by Rees et al. (1998) who found higher RT in SIM on the TOMM. They concluded that simulators had to recognize the correct answer before providing the wrong answer, which increases RT. If this is so, RT may be sensitive to feigning on easy tasks (i.e., tests such as the TOMM) but may be less so on tasks with a wider range of difficulty (e.g., PDRT). More research on the utility of RT is warranted due to the variability in samples and tests used.

Another surprising result regarding RT was noted in Section 1. RT was lower in SIM than in NC. One possible explanation for this is that SIM did not try as hard (i.e.,

spend the time to consider response alternatives) as NC to answer correctly. This result may be due to limited understanding of their instructions, poor motivation to feign, or due to an attempt simply to miss items without much thought. If the latter is true, then SIM may have underestimated the importance of RT and over-relied on the number incorrect as an indication of impairment. The fact that NC, though slower in RT, answered significantly more items correctly than SIM on Section 1 supports this view.

Response Time X Performance Curve. Wogar, Van den Broek, Bradshaw, and Szabaldi (1998) used a forced-choice recognition memory paradigm and discovered RT X PC interactions in simulators of neurocognitive impairment. They found a disproportionate increase in latency in simulators as item difficulty increased, suggesting that simulators misjudged item difficulty and overestimated the time required to answer.

Binks et al. (1997) used RT to detect feigning on the Dot Counting Test, and identified several indices of varying efficacy. Though RT by itself offered little to feigning detection, the number of times RT for grouped dot counting (i.e., easier task) exceeded RT for ungrouped dots (i.e., more difficult task) was an effective discriminator. Like Wogar et al. (1998), this strategy capitalizes on the fact that RT increases with item difficulty.

These two studies suggest that much of the value of the RT strategy lies in the fact that, when feigners misjudge the difficulty of an item as more difficult than it actually is, they likely increase response latencies. Alternatively, they may pause long enough to debate whether they should answer correctly, which increases RT (Rees et al., 1998).

RT may work best on easier tasks. However, on the TOCA it appeared that even easy items were too difficult for CL to answer more quickly than SIM. Even those items that 95% of CL answered correctly took much longer for CL than SIM to answer. Alternatively, it is possible that SIM did not appreciate the impact brain injury has on mental speed and, despite often slowing RT, still responded quicker than CL. Support for the profound effect of brain injury on RT is found in the fact that mTBI patients, though getting significantly more items correct, did not respond more quickly than many seriously impaired CL, $t(42) = 0.99$, $p = .33$.

The examination of RT in feigning detection is a relatively new endeavor. It appears to have considerable promise, particularly when used with other strategies, but requires more study. Several supplemental analyses were conducted in the current study for this reason.

RT on Easy Items. This RT scale measured RT on only easy items (i.e., 13 items answered correctly by 90% of CL). As mentioned previously, the debriefing revealed that some SIM slowed their RT to appear brain damaged. Because SIM were unaware of the relative difficulty of any particular item, it was hypothesized that they may overestimate the time needed to respond to easy questions. To assess this possibility, SIM RTs were compared to NC and CL RTs on the most effective Floor Effect items (FE-90%). Contrary to expectations, CL responded slower on easy items than SIM.

RT for mTBI Patients. The RTs of the 11 participants who had sustained mTBI were compared to SIM RTs. Although statistical comparisons including mTBI were limited due to the small sample size, SIMs RT were substantially lower than mTBI RT.

This result supports the conclusion that even the mildly impaired patients in this sample responded significantly slower than SIM.

Symptom Validity Testing

Although only two participants scored significantly below chance on the TOCA, many participants scored at chance at some time (usually in Section 2). A likely explanation is that some items are so difficult as to result in chance performance on these items. Alternatively, some incorrect items may be more likely to be chosen than others because they appear superficially correct. A preliminary individual response analysis performed on the incorrect responses most likely to be chosen did not reveal common characteristics or patterns. For example, the position of the response (i.e., whether it was first, last, or in between) appeared to be unrelated to likelihood of selection. This observation was noted for both CL and SIM. The research employing SVT may identify characteristics of these incorrect responses that make them more likely to be chosen.

Stimulated by Pankratz's (1979) and Binder's (1993) investigations, more recent research on SVT has yielded varying results. In general, successful use of SVT has been documented in case studies (Frederick, Carter, & Powel, 1995; Pankratz, 1979; Wiggins & Brandt, 1988), whereas less success has been noted in experimental designs (Amin & Prigatano, 1993; Beetar & Williams, 1994; Frederick & Foster, 1991; Guilmette & Hart, 1993; Hiscock & Hiscock, 1989; Rees et al., 1998; Tombaugh, 1997). This finding reflects the excellent specificity but poor sensitivity of SVT. It also helps explain why case studies have typically been used to espouse the use of SVT. Researchers have been

able to confidently classify an individual as feigning due to the unequivocal specificity of SVT, but have not been able to identify groups with SVT due to poor sensitivity.

All the reports reviewed for this study employed a 2-alternative forced-choice format, therefore chance performance was set at 50%. In contrast, this study employed a 4-alternative format; items required 2 answers to be selected from 4 alternatives. This use of SVT is somewhat more complex than the 2-alternative version. As a result, it may be more difficult for the feigner to determine the exact level of chance performance and alter performance accordingly.

Two simulators were detected using the Truncated SVT scale, while none were detected using the other versions. This finding suggests that removing items that are too difficult to answer correctly and the items too easy to miss may improve the clinical utility of the SVT strategy. Chance likelihoods of .08 and .25 appear to be far too low to detect the vast majority of individuals asked to feign neurocognitive impairment.

In sum, the results from the SVT strategy support previous reports that it is highly useful when a client performs significantly below chance. Though SVT correctly classified only two individuals as feigning in the current study, no other plausible explanation is found for their performance. Its specificity is an unparalleled strength of the strategy. Nonetheless, the SVT strategy on the TOCA failed to identify the majority of simulators as such and yielded very poor classification rates.

Overlearned Material and Section 3 of the TOCA

Very recent research has demonstrated that overlearned verbal material can differentiate feigned impairment from honest patients. Boone et al. (2000) posited that

because the general public does not seem to know that overlearned information is relatively spared following brain injury, they tend to answer these items incorrectly more often than brain-injured patients. Based on these findings, Boone et al. found that their test of letter recognition yielded a sensitivity of 76.5% and specificity of 85.1%. The similar scores in Section 3 of this study suggest that it shares the same principle of overlearned information.

Section 3 of the TOCA is a sentence-stem completion task. The questions utilize highly recognizable and overlearned phrases that most individuals learn in childhood (e.g., nursery rhymes). As a result, most of these items are considered to be quite easy and insensitive to neurological compromise. Section 3 score may represent a latent Floor Effect strategy due to its ease relative to the other 2 sections.

Measures Employing Multiple Strategies

Very few studies have employed multiple strategies (e.g., combining Floor Effect and MoE strategies). Frederick and Foster's (1991) modification of the TONI, the VIP (Frederick & Foster, 1997), and CARB (Allen et al., 1997) are three measures that utilize more than one strategy. Table 41 displays the measures, strategies employed, item content, and the neurocognitive domains covered in the tasks. Interestingly, 2 of the 3 measures plus the TOCA are computer administered. This commonality likely reflects the fact that computers facilitate the use of complex decision rules and the precise recording of RT.

The current results are consistent with early work on the TONI (Frederick & Foster, 1991). Briefly, they used a mathematical product of Performance Curve and

Table 41

Detection Strategies, Item Content, and Neurocognitive Domains Assessed by Four Multi-scale Measures of Feigning: The CARB, TOCA, TONI, and VIP

Measure	Strategies Employed	Item Content	Neurocognitive Domain[a]
CARB	Floor Effect RT	Verbal (digits)	Recognition Memory
TOCA	Floor Effect Magnitude of Error Performance Curve Response Time SVT	Verbal & Nonverbal	Verbal comprehension Visuospatial analysis Nonverbal reasoning Working memory
TONI	Performance Curve (PC) Response Consistency (RC) PC X RC Floor Effect	Nonverbal	Nonverbal reasoning
VIP	Performance Curve (PC) Response Consistency (RC) RC X PC Floor Effect	Verbal& Nonverbal	Word knowledge Nonverbal reasoning

[a] The domains listed for the TOCA and the VIP are still under investigation.

Consistency Ratio on a modification of the TONI (Brown, Sherbenou, & Johnson, 1982) to differentiate simulators from patients. Their later modification of the TONI, the VIP, expanded classification possibilities from 2 (feigning and honest) to 4 (irrelevant, careless, compliant, and malingered). This change makes it difficult to directly compare the studies. Nonetheless, it appears that PC and Response Consistency combinations are effective in a nonverbal context. Classification accuracy dropped when a verbal test was added to the TONI (see the VIP; Frederick & Foster, 1997). This finding is consistent with Van Gorp et al.'s (1999) finding that suspected malingerers scored

disproportionately lower on nonverbal tests. They suggest that feigners are less familiar with the effects of brain injury on nonverbal abilities and are less effective at feigning impairment on these tasks as a result.

In contrast, the data from the TOCA suggest that Floor Effect and MoE in combination work particularly well on verbal tasks. However, it appears that the relative ease of items in the verbal task likely contributed to the success of the strategies.

The CARB (Allen et al., 1997) also utilizes more than one strategy to measure feigning, namely Floor Effect and RT. The authors reported that simulators miss more easy items than patients and RT was slower in simulators than in patients. Though the classification rates were quite good, the false positive rate was too high. The current study yielded a superior combined-scales classification rate as compared to the CARB. The data from the CARB suggest that the Floor Effect by RT interaction is clinically useful, though more research to replicate the findings is needed.

In summary, the preceding review lends support for the use of conceptually-driven detection strategies in feigning detection. Prior research tended to focus on empirical evidence but neglected the underlying constructs of proposed techniques. This neglect likely resulted in less effective strategies and research. Strategic feigning detection appears to hold promise for scales used both individually and in combination.

Methodological Issues Related to Simulation Design

Research on feigning detection in clinical neuropsychology proceeds despite the lack of a gold standard. It is therefore important to review the methodological constraints in feigning detection research. The primary limitation to the majority of feigning studies

is questionable external validity. Several means of increasing external validity exist but relatively few studies have addressed them. The following section summarizes the potential effects of instructional sets, incentives, coaching, and manipulation checks on external validity.

Threats to External Validity

Investigators of feigning must choose from three research designs: known-groups, simulation, and differential prevalence. Although several case studies describe the neuropsychological profile of a known malingerer, no published experiments are available that employ a pure known-groups design. Recently, the number of experiments using "suspected malingerers" has increased, with the assumption that external validity is improved in these studies. In this design, participants who meet certain criteria are classified as suspected malingerers. These criteria typically include persons who meet two or more of the following criteria: (a) fail one or more specific tests of motivation, (b) are pursuing litigation, (c) produce inconsistent results on neuropsychological evaluation, and (d) have highly unusual complaints (see Greiffenstein et al., 1994 for a discussion of criteria). This group is compared to simulators and controls. While the requirement of meeting more than one of the above criteria reduces false positives, it is not a fail-safe method, thus the term "suspected malingering" should not be taken to mean "malingering."

The majority of research continues to employ the simulation design. A strength of the simulation design is that it allows the investigator to increase internal validity by controlling for sources of error. The main drawback is that the generalizability of the

findings to actual feigners is unknown. Several common components of simulation design affect the utility of the results.

Instructional Set. Detailed instructional sets are crucial to good simulation designs. Despite this need, many feigning studies provide brief and nonspecific instructions. Nies and Sweet (1994) listed, "vague instructions to simulators" as a weakness for all of the 19 simulation studies they reviewed.

Several elements must be considered with respect to instructional sets (Rogers, 1997):

- First, the instructions should be comprehensible. In other words, the simulators should understand their role and what they are being asked to do.
- Second, the instructions should be specific; general instructions may not adequately elicit the response style under investigation.
- Third, the instructions should provide a scenario or context to which the simulators can relate. Rogers and Cruise (1998) found that contexts that simulators found unfamiliar or irrelevant alienated them from their role.
- Fourth, the instructions should attempt to capture the potentially harmful consequences of being caught faking. This is a challenging but important goal for external validity because many real-world malingering scenarios involve potentially devastating consequences if caught malingering.
- Fifth, the instructions should emphasize the importance of credible responding. Research has shown that failure to do so can result in blatantly obvious feigning (Franzen & Martin, 1996).

The specific effects of instructing participants to feign a non-existent injury as opposed to instructing them to feign exaggerated symptomatology of a real injury have not been examined. Instructional sets that emphasize exaggeration in order "to get what you really do deserve" may yield different patterns of responding than those that emphasize the "take advantage of the system while you can" mentality.

The current study allowed as much time as participants felt they needed to read the instructions and to prepare for their role. No participant used more than 10 minutes. The simulators tended not to use their time to strategize, rather they tended to review the instructions. How this compares with other studies is largely unknown; many studies do not comment on the time allotted to simulators to prepare or what they did during that time. Both of these factors still require investigation (Rogers et al., 1993).

Incentives. Incentives are considered an important part of simulation design because they represent an attempt to approximate real-world influences on client behavior. Rogers (1997) suggested that four elements are instrumental to the effective use of incentives in simulation research:

- First, the magnitude of the incentive should be adequately representative of either favorable or unfavorable consequences for the simulator. The incentive may be monetary or may be symbolic of success (e.g., being identified as the "winner" or "best"). However, research has not yet been conducted on whether being identified as the most proficient simulator is actually rewarding. Less pejorative labels (e.g., best actor) may be more appealing to simulators.

- Second, the type of incentive also must be considered. In college populations, course credit may be considered as rewarding as nominal monetary incentive. However, course credit may not be as representative of real-world incentives as money.
- Third, the probability of receiving the incentive may affect dissimulation. Some experiments offer nominal rewards to all participants while others offer rewards only to those feign successfully. No data exists on which approach is more likely to enhance credibility.
- Fourth, few studies have examined the effect of negative incentive. The work of Rogers and Cruise (1998) suggests that negative incentives are effective and increase external validity of simulation design. Specifically, when feigning depression, participants with negative incentive were more focused in their feigning and produced more symptoms of depression, while reporting fewer symptoms unrelated to depression.

Studies have yielded mixed results regarding the effect of incentive on simulators; some finding differences (Martin et al., 1993) and others not (Bernard, 1990; Schretlen & Arkowitz, 1990). Coleman et al. (1998) suggested that these results are due to the lack of a standard for how motivation is defined and assessed in an experimental paradigm. Moreover, they argued that conclusions about the effects of motivation are problematic because the incentives do not approximate real-world incentives. In contrast, Rogers and Cruise (1998) suggested that incentive can be manipulated and differences in performance do occur based on the type of incentive. In sum, the variability of the

findings regarding the use of incentives to enhance feigning underscores the importance of clearly documenting the type and amount of incentive used.

The impact of the type of incentive appears to be partially dependent upon the accompanying instructional set. For example, without instructions to feign believably, simulators may overplay their role to obtain the incentive and consequently reduce classification rates (Franzen & Martin, 1996). No data exist in the clinical neuropsychology literature regarding the effect of negative incentive in simulation design.

Coaching. Coaching involves providing the patient with information about the disorder in question or about the strategies likely to be used to detect feigning. Coaching techniques, especially sophisticated coaching, can seriously diminish the clinician's ability to differentiate feigned from honest performance.

The effect of coaching on the performance of simulators has received increased attention in recent years. This may be due to psychologists increasing awareness that some attorneys feel obligated to coach as part of their duty to zealously represent their clients. Thus, especially in forensic contexts, the possibility of coached dissimulation must be considered.

Research suggests that naïve simulators are easier to detect as faking than sophisticated (i.e., coached) simulators (Franzen & Martin, 1996; Martin et al., 1993). Specifically, it appears that receiving information about the detection strategy itself rather than the disorder in question improves simulation (Rogers, 1997).

Questions have been raised about the ethical implications of publishing instructions that include hints or coaching because those who are coached may use the information for fraudulent purposes in the future (Sweet, 1999). Coaching may also jeopardize test security and validity if clients obtain important information regarding the specific questions in the exam. One solution suggested by Rogers (1997) is the use of intricate strategies. This approach involves informing the patient that multiple detection strategies will be employed during the assessment. This information requires the would-be feigner to consider multiple detection strategies while attempting to produce a believable profile. This process of warning minimizes concerns about the beneficial effects of coaching on the feigner because attempts to feign are potentially thwarted, rather than facilitated, by the warning. Theoretically, any number of warnings will not improve their chances of going undetected (if they try to feign) because it is too difficult to simultaneously entertain the strategies and the desired diagnostic profile.

The effect of warning patients that feigning detection measures will be used during the evaluation has been scrutinized recently. Youngjohn, Lees-Haley, and Binder (1999) argue that warning malingerers increases their sophistication rather than reducing malingering behavior; thus, they conclude that warning clients is ill-advised. In contrast, Johnson and Lesniak-Karpiak (1997) argue that warning helps reduce the potential for feigning by informing the clients that they probably will be caught if the feign. They also contend that using detection techniques without informing the patient may be viewed as deceptive and therefore ethically questionable. Further investigation of the effect of

requiring individuals to entertain multiple decision-trees while performing tasks of varying difficulty appears warranted.

Rogers (1997) has argued that the caution on the TOCA serves as a complex distractor because it informs the respondent that multiple and simultaneous processing is necessary to foil the test. Because test-takers are presented with an unwieldy number of ways they may be detected, if they choose to malinger they will be overloaded by the simultaneous tasks of feigning credibly and avoiding several detection strategies. As a result, the likelihood of detection increases. Conversely, if they decide that it is too difficult to feign because of the multiple detection strategies, they will perform honestly. Either way, crucial information is provided.

This study provides some evidence that warning simulators of the presence of detection strategies may alter their performance on the TOCA, relative to non-cautioned simulators. Cautioned simulators were slightly easier to detect than non-cautioned simulators. The number and complexity of scales on the TOCA may have proved too overwhelming and decreased the advantage of being cautioned that detection strategies were in place. Consequently, cautioned simulators did not benefit from the warning.

Manipulation Check and Debriefing. The primary purpose of the manipulation check is to exclude those participants who did not understand or comply with the instructions. The current study's manipulation check resulted in the exclusion of nine simulators who did not adhere to the instructions. Their exclusion was necessary to ensure the accuracy of experimental conditions. Despite their importance to simulation research, Nies and Sweet (1994) found that very few analogue studies included a

manipulation check. This oversight suggests that most of the published simulation studies may include participants who do not satisfy experimental conditions, which calls into question their clinical utility.

The manipulation check investigated the degree to which simulators perceived themselves as successful at fooling the TOCA. Only about one-fourth of the SIM thought they were successful. Almost all of these participants were NCS, suggesting that the caution may have significantly reduced CS's confidence in feigning without detection. This result likely reflects the fact that NCS were not made aware of the detection strategies being used. In contrast, CS were aware of the multiple detection strategies and may have assumed that they were unsuccessful. This conclusion may also explain why some cautioned simulators feigned more obviously than non-cautioned simulators. They may have felt it too unlikely to go undetected in light of all the detection strategies listed in the caution.

Participants' estimations of how well they faked were poor predictors of actual success. Those who rated themselves as successful were no more effective at avoiding detection than those who felt they were unsuccessful. For example, no difference was found between successful and unsuccessful participants on Section Score (see Results section). Likewise, their judgments of success were unrelated to actual performance on the FE-95% and 7.1%-MoE3 scales, $\underline{t}(49) = 1.45$, $\underline{p} = .15$ and $\underline{t}(49) = 1.02$, $\underline{p} = .31$, respectively.

Implications for Clinical Practice

The following section reviews the practical implications of the current findings in light of previous research. The clinical utility of each strategy is summarized followed by a discussion of the potential of combining strategies. The section concludes with a review of ongoing problems of malingering classification.

Magnitude of Error

The TOCA's operationalization of the MoE strategy yielded excellent classification rates. A clinical strength is that clients may be unaware of its utility (Tenhula & Sweet, 1996). In addition, their chances of consistently identifying the particular incorrect response are poor, even if they are aware of the strategy.

MoE is promising for use on existing clinical measures. In particular, multiple choice measures commonly used in neuropsychological batteries, such as the Wisconsin Card Sort Test (WCST; Berg, 1948), could be adapted to this procedure (Rogers et al., 1993). For example, a client who chooses incorrect responses that fewer than 10% of a clinical comparison group chose is likely to indicate feigning. The number of "other" responses may hold promise as a measure of Magnitude of Error because it is rarely encountered in clinical practice.

The work of Martin et al. (1998) suggests that clinicians may also employ MoE on the WMS-R (Wechsler, 1987). This application has appeal clinically because the examiner gets information regarding visual and verbal memory from a well-known measure, as well as information regarding feigning. The potential limitation is that this application has yet to be cross-validated.

In summary, the MoE strategy is conducive to implementation on a variety of tests. It is also likely difficult to determine the responses used in the scales, even in the face of coaching. Thus, the MoE strategy continues to hold promise in feigning detection, particularly on multiple choice measures. Nonetheless, given the lack of other validation studies, the MoE strategy requires further study before clinicians can confidently use this strategy in professional practice.

Floor Effect

The Floor Effect strategy has numerous advantages and has received extensive attention in the neuropsychological literature for several practical reasons (see Nies & Sweet, 1994 for a review). The Floor Effect is appealing to practitioners because it is conceptually simple and can be applied in a variety of ways to existing measures. Also, the concept of the Floor Effect does not require broad knowledge of malingering research or learning theory. Moreover, most tests employing the Floor Effect are brief and therefore appealing to practitioners who often face time pressures. Finally, the most common tests using the Floor Effect strategy are relatively inexpensive and portable (e.g., FIT; TOMM).

Clinicians should also be aware of the potential limitations of the Floor Effect strategy in practice. Tests that employ the Floor Effect strategy are subject to potential face validity concerns. It may be clear to a sophisticated malingerer, for example, that the 15-Item Test (FIT) is very easy. Consequently, the malingerer may perform within expectations on this measure and perform poorly on more difficult tasks. Therefore, the order of test administration is also a concern. If a simple recognition task is administered

after more difficult measures of memory, the malingerer may realize that it is too easy relative to the other measures. This transparency is less problematic on measures of ability because they are face valid as measures of ability. These measures were not originally designed to detect feigning and they utilize more recently developed (and unnoticeable) scales of poor effort/motivation (e.g., WAIS-III; CVLT; Category Test). Measures such as the VIP (Frederick & Foster, 1997) and the TOCA (Rogers et al., 1996) embed Floor Effect items with items of varying difficulty, making it more difficult for the would-be feigner to anticipate which items to miss.

At least one use of the Floor Effect strategy raises potential ethical concerns. Clinicians are prompted to emphasize the difficulty of tests such as the 15-Item Test. Rogers et al. (1993) postulated that this practice is ethically questionable because the examiner is engaging in the same behavior considered inappropriate in the client.

In summary, the Floor Effect strategy works relatively well and has been well researched. It remains a very promising area of research and is often favored by clinicians due to its brevity and simplicity. Clinicians should feel sufficiently confident in using most current tests that employ the Floor Effect as a screening measure. If the results of the screen suggest feigning, additional tests of feigning should be administered. However, it typically has been used to assess feigning of only one ability (e.g., memory) and it is limited by potential face validity concerns.

Response Time

Much of the clinical appeal for the RT strategy lies in its simplicity and availability; many existing measures employ time limits as a part of the assessment of

neurocognitive function. However, the RT strategy's utility appears to depend, at least in part, on the characteristics of the test being used (e.g., item difficulty and length). RT may work best on tasks with items of varying difficulty; RT on easy items could be compared to RT on difficult items. Discrepancies in RT (e.g., shorter RTs on more difficult items) may indicate feigning. A potential weakness of RT is that it is sensitive to bona-fide brain dysfunction, particularly attention deficits. Clinicians must be very careful when making determinations about feigning status and should not rely on RT in isolation. Research is yet to provide unequivocal evidence as to its utility, but given recent data, RT likely holds considerable promise as a measure of neurocognitive feigning.

Performance Curve

The clinical appeal of the PC strategy may lie in its subtlety and complexity. With respect to subtlety, Tenhula and Sweet (1996) reported that only a very small percentage of simulators considered the difficulty of performance to be an important aspect for detection. Regarding complexity, even if the simulator is aware that the PC strategy is being used, its complexity makes it very difficult to appear impaired and avoid detection. These findings coupled with the PC strategy's applicability to existing measures of ability, likely make it particularly appealing to clinicians. A potential limitation to PC is that items of varying difficulty are necessary, which may lengthen the test. It also requires extensive statistical analyses to develop the scales and determine the levels of difficulty. Without computer-based scoring, clinicians cannot easily apply the strategy to existing measures.

Clinicians can be quite confident that a client has intentionally altered his or her responses if there is an atypical performance curve. For instance, no plausible alternative explanation can be established for a positive curve as items increase in difficulty.

Symptom Validity Testing

The results of the current study suggest that SVT has limited clinical value. As stated previously by other investigators (e.g., Pankratz & Binder, 1997; Sweet, 1999), the SVT strategy's main drawback is its poor sensitivity. While a positive finding using SVT is highly specific to feigning, negative findings do not rule out feigning. Although several participants scored at chance at some point on the TOCA, only two participants performed significantly below chance.

Importance of Multiple Measures of Feigning

According to Rogers (1997) and others (e.g., Frederick & Foster, 1991; Sweet, 2000), feigning detection is facilitated by employing detection scales in more than one neurocognitive domain. It appears that it is more difficult to sustain a credible profile of feigned deficits than to feign just one deficit. Also, these tests likely will miss feigning of a deficit other than the one being assessed (e.g., memory). These observations underscores a drawback of measures of feigning of a single ability.

Potential malingerers likely will find it difficult to anticipate the multiple expected patterns of performance, particularly regarding the more obscure aspects of the test itself. For example, Bernard, McGrath, and Houston (1996) suggest that certain indicators on the WCST can effectively differentiate simulators from patients. They found that simulating malingerers were unaware of the concept of Perseverative Errors

(subtle indicator), and consequently failed to make the same proportion of Perseverative Errors-to-Categories Completed that normals and patients make. They concluded that although malingerers may be able to track Categories Completed (an obvious indicator), they fail to recognize the need to track performance on less apparent indexes. As a result, both naïve and coached feigning were vulnerable to detection.

A potential for concern for clinicians is that tests that include a wide range of item difficulty and assess more than one ability tend to be longer than tests that employ one strategy or that measure one neurocognitive domain. Few tests have been designed specifically to measure both feigning and ability, largely because many clinicians value brief tests of feigning. Thus, a potential drawback of these measures, including the TOCA, is administration time. However, to maximize the usefulness of data obtained in this time, the TOCA attempts to provide sensitive and specific indices of feigning status as well as estimates of working memory, nonverbal reasoning, and verbal comprehension.

In sum, the data from past and present research suggest that clinicians should use multiple tests of feigning. These tests should be well researched and correctly classify simulators at least 75% of the time. In other words, they should fall within the Probable range of certitude (Rogers, 1997). In addition, clinicians should assess feigning in more than one domain whenever possible.

Limitations to Single Sources of Information

In evaluating feigning status, it is essential that clinicians use data derived from tests of feigning in conjunction with other clinical data. Sweet et al. (2000) have

discussed the limitations to single sources and essentially argue for a multidimensional threshold model of feigning detection. Rogers (1997) reminds clinicians to consider carefully the strength and consistency of findings from all parts of the assessment, the presence of other explanations, the degree of malingering if present, and consequences of categorizing a patient as malingering.

In sum, the data obtained from feigning tests are not by themselves indicative of malingering. The clinician is ultimately responsible for ruling out the other possible explanations for the performance. For example, the only criterion separating malingering from Factitious Disorder is the source of the motivation (internal or external). Thus, it is particularly important that clinicians understand the context and the potential for gain. A patient who performs below chance on SVT but is without identifiable external gain should not be classified as malingering. Clinicians are also encouraged to remember that malingering and genuine impairment are not mutually exclusive (Rogers, 1997).

Clinical Decision Models and Diagnostic Criteria for Malingering

Pankratz and Binder (1997) proposed both a threshold model and guidelines for a clinical decision model. They suggest that the threshold for suspecting dissimulation should be set relatively low so as to ensure adequate sensitivity to the presence of feigning. If any one of several criteria is present, clinicians are urged to intensify their examination of the client. Specifically, if the threshold is met, the clinician is to investigate further the client's social history, employment and litigation status, medical and psychiatric history, and other contextual issues as part of the clinical decision-making process. Though they recommend obtaining empirical evidence of feigning, the

determination of malingering is multifactorial and is essentially a matter of clinical judgment.

Threshold and clinical decision models are of limited value without an explicit definition of malingering and specific classificatory criteria. Slick, Sherman, and Iverson (1999) noted that although Rogers (1990b) has proposed specific criteria for malingered psychopathology, a comparable definition with specific reference to neuropsychological malingering does not exist. Slick et al. established their diagnostic criteria for malingered neurocognitive dysfunction based on the multi-tiered model used for the diagnosis of Alzheimer's Disease. Specifically, classifications of Definite, Probable, and Possible malingering were defined. This model allows for the fact that it is not often possible to identify malingering with absolute certainty. Each definition differs according to the level of certitude based on the evidence of malingering (i.e., number of criteria met).

They contend that their criteria redress the major limitations to prior guidelines by specifying: the definition of malingering in neuropsychology, unambiguous diagnostic criteria, the relative weighting of each criterion, the role of clinical judgment, and levels of certainty. However, their criteria for Definite Malingering is likely too restrictive; Definite Malingering cannot be diagnosed without significantly below chance SVT data. Likewise, criteria for Possible Malingering does not adequately differentiate it from other disorders involving altered response styles. Nonetheless, their work is important to clinicians because it represents the first systematic attempt to establish specific criteria of for malingering in the neuropsychological context.

Other Implications for Practice

Much of the initial research on feigning detection in neuropsychology involved tests of verbal attention and memory. For example, the HDMT (Hiscock & Hiscock, 1989), and the PDRT (Binder, 1993) are recognition measures that ostensibly assess immediate memory of strings of verbal information (digits). Verbal learning and memory tests were also used to develop feigning detection scales (e.g., RAVLT, CVLT, WMS-R Logical Memory subtest). The "b test" (Boone et al., 2000) and Section 3 of the TOCA are verbal tasks that do not have a substantial immediate recall component that have been shown to differentiate simulators from patients. However, in recent years, the literature has also demonstrated the effectiveness of feigning detection scales on nonverbal measures. The TOMM (Tombaugh, 1997), DCT (Binks et al.,1997), WCST (Bernard et al., 1996). In sum, clinicians are no longer limited to recognition tests of verbal stimuli; clinicians can evaluate feigning in both verbal and nonverbal domains.

The multiple-choice paradigm continues to be effective in feigning detection. The majority of studies in this review utilized this format. Multiple-choice testing is conducive to feigning detection because it is administratively simple, easily formatted to a variety of measures (e.g., recognition portion of memory tests), and provides objective findings. However, tests of feigning that do not employ multiple-choice paradigm also appear be clinically useful. Most of these involve comparisons to expected performance patterns that the client is unaware of (e.g., the fact that recognition memory relative to free recall is insensitive to brain injury).

Computer administration of neuropsychological tests, including tests of feigning, is increasingly common. A major advantage of the VIP (Frederick & Foster, 1991) is that it is computer administered. Much of the test's discriminatory power is established through complex statistics and logarithms that would take an enormous amount of time to calculate by hand. The computer scoring program rapidly calculates these statistics and generates easy-to-read reports and tables. A potential disadvantage of computer-administered tests is their limited availability compared to paper and pencil tests and the increased cost of software. Some monitors and keyboards, especially on laptops, may be difficult for patients with even mild visual problems to use. The same difficulty exists for clients with psychomotor impairments.

Strengths of the Current Study

Most of the strengths of the current study involve attempts to reduce threats to external validity. Consistent with Rogers (1997), several methodological considerations were incorporated during the development of this study.

Incentive. The implementation of both positive and negative incentives is considered a strength of this study, in light of the paucity of studies that used both. The positive incentive was provided contingent upon successful feigning (see Bernard, 1990). This stipulation is intended to avoid the "extreme feigning" noted by Franzen and Martin (1996). Simulators were informed that course credit was contingent upon success as compared to a standard criterion, rather than contingent solely on their performance as compared to the other participants. Thus, it was hoped that the incentive would be motivating to both competitive-natured and less competitive participants. Unfortunately,

because only the top three feigners were rewarded with $20.00 each, some simulators felt it too unlikely for them to obtain the reward.

Another strength of the study was the inclusion of negative incentive (publicly posting the names of the three worst feigners). Rogers and Cruise (1998) have shown that potential embarrassment can be used to approximate the negative real-world consequences of being caught faking depression. More research on the effect and type of negative incentives in clinical neuropsychology is needed.

Instructional Set. The instructional set used in this study was adapted from that used by Rose et al. (1998). It included both an example of how to be credible (e.g., not being too obvious when faking) and some specific results of head injury (e.g., inattention; slowed processing speed). The scenario also provided a "reminder" of the potentially negative effects of being caught faking (e.g., jail), thereby potentially strengthening the effect of negative incentive. All respondents reported that they had little difficulty understanding their role.

Manipulation Check. Nies and Sweet (1994) reported that only two simulation studies in their comprehensive review of the literature included a debriefing or a manipulation check. None of 19 studies provided highly specific instructions to simulators. This study included a manipulation check of simulators' feigning strategies, understanding of the instructions, prior knowledge of brain function, perceived success, and effort. Nine participants who did not attend adequately to the instructions were excluded from analyses.

Influence of Ethnicity on Likelihood of Detection. The examination of the effect of ethnicity on vulnerability to detection represents a strength of this study. Prior research has not assessed this relationship. The present study found that ethnicity did not appreciably influence likelihood of being detected as feigning, with one exception. Among simulators, Hispanic Americans were more likely to produce lower scores on the FE-95%-I scale, which made them more likely than Caucasian Americans to be correctly classified as feigning. Though preliminary, this result suggests that Floor Effect scales may be sensitive to cultural influences. Research is needed to confirm these findings and to explore other potential influences on feigning ability and detection. The implications of this finding are discussed further in Appendix K.

Limitations to the Current Study

Limitations to the current study focus mainly on the related issues of sample selection and simulation design. For example, although large enough to conduct null hypothesis testing and DFAs, the overall sample size is not sufficiently large for cross-validation. Also, the potentially spurious findings from multiple comparisons will be discussed. This section concludes with a discussion of the apparently modest effect of providing incentives.

Clinical Sample. The current study utilized a heterogeneous clinical sample was selected in an attempt to increase external validity. However, its diversity may have obscured differences between groups and diminished classificatory accuracy of the TOCA. A more homogenous clinical sample (e.g., all mTBI patients) may have yielded

more homogeneous performances. The current study also included a small number of mTBI patients, which limited analysis of this group.

Simulating Sample. The current design was vulnerable to problems of generalizability, despite the steps taken to increase external validity. For example, the representativeness of college students as actual malingerers can be questioned (Nies & Sweet, 1994). College students may not be as motivated by course credit, curiosity, and nominal rewards, as bona-fide malingerers who may be motivated by much larger consequences. The fact that some simulators did their best, despite being asked to adhere to instructions to the contrary, lends support to this observation.

Incentives. The adequacy of the incentives can be questioned. Although the inclusion of negative and positive incentive was considered a strength of the study, the reports of SIM indicate that it may have been insufficient. Some felt that it was very unlikely that they could be one of the three "best."

For simulation research to increase external validity, meaningful and "real-world" incentives are needed. A potential dilemma arises, however. If incentives offered in simulation research can match those in the real-world (which is highly unlikely), at what point does incentive become coercion? Simulators themselves may shed light on this subject by responding to questions in future debriefings about the relevance of the rewards to them.

The study focused on theory-based detection strategies as a method of limiting the number of analyses. Still, the present study did involve numerous comparisons. Several detection scales were developed based on five alternative methods of categorization, each

of which was involved in multiple comparisons. In hypothesis testing, a large number of comparisons increases the probability of making a Type I Error (i.e, false positive error). In addition, although most comparisons were made between just two groups, at times four groups required investigation.

The current sample consisted of patients with head injury or cerebrovascular accident. Research conducted with psychiatric samples is needed to help establish the effect of psychiatric illness on neurocognitive performance. Likewise, studies of the effect of asking head-injured patients to feign injury are needed as they may reveal different patterns of performance.

Directions for Future Research

Prevalence Rates and Utility Estimates

Current prevalence estimates are highly variable both within and across settings (Gouvier, 1999; Rogers, 1997). The problems inherent to estimating base rates are difficult to overcome and appear to a part of a vicious cycle. Malingerers are extremely unlikely to identify themselves as such and clear indications of the principle motivation to feign are rare. These issues make differential diagnosis difficult, which in turn makes the prevalence rate difficult to estimate. It is recommended that research continue to include both estimates that take into account base rates and those that do not. Reports of classificatory accuracy should include estimates based on the current study's base rate of malingering as well as a more likely base rate. This second estimate may yield more accurately real life applications of the research findings. Alternatively, estimates based on +/- 1 SD from the base rate may be useful.

The establishment of accurate prevalence rates of malingering is complicated by the largely unknown prevalence rate of bona-fide patients who also malinger. The growing literature on persisting post-concussive syndrome suggests that many patients with a mild brain-injury may also over-report cognitive or psychological symptomatology at 3 and 6 months post-injury (Ponsford et al., 2000; Ruff, Camezuli, & Mueller, 1996). Whether this is commonly labeled as malingering is unknown and deserves investigation.

Lees-Haley et al. (1996) suggest that malingering is not adequately addressed in neuropsychological evaluations and argue that clinicians underestimate the occurrence of malingering. They contend that a disproportionate reliance on clinical judgment, rather than test data, has resulted in excessive false negatives. Consequently, current prevalence rates of malingering may be underestimates. Given the large amount of research that has been published since Lees-Haley et al.'s survey, attitudes about malingering may have changed. A follow-up survey of current attitudes and practices appears warranted.

Computer Technology

Recent advances in computer technology may hold particular promise in test development and administration. Computers allow for complex presentations of stimuli and for analyses to be performed with virtually no variance in presentation. Also, the benefits of computer-based statistics programs are widely known and have made formerly unwieldy procedures quite easy. These advancements have increased researchers' ability to assess diagnostic accuracy via complex statistics such as DFA.

The accessibility to the internet and to other scientist-practitioners via email may open the door to accumulating prevalence data. Clinicians by themselves likely do not see

enough known malingerers to form known-group studies. If data are pooled from many clinicians, prevalence rates, and important sociodemographic and psychosocial characteristics of those who malinger may be revealed.

Malingering Profile

Few validated patterns of feigned performance across neuropsychological tests have been discovered to date. Highly similar demographic and personality characteristics among feigners have been even less forthcoming. Future studies may benefit from investigating characteristics common to individuals who score similarly on neurocognitive feigning tests. For example, what are the similarities in sociodemographics, personality, reasons for referral, and neurocognitive functioning between those who score relatively high on a given scale versus those who score low? These features may lead to specific profiles to which clinicians can compare when assessing neurocognitive feigning.

Malingerers appear to be a very heterogeneous population. Classification of malingerers may be more efficient if the scope is narrowed. For example, Rosenfeld, Sands, and Van Gorp (2000) suggested that tests of feigning should focus on individual characteristics or settings, rather than trying to develop tests that detect all feigners in all settings. They cite Schretlen et al. (1991) as a good example of the problems with using too large of a scope. Schretlen et al. found that a large percentage of patients with genuine amnestic disorder scored below the recommended cutting score for feigning on the FIT. Given the previous work suggesting that the FIT is a useful screen, they

concluded that the FIT may not be a valid tool for assessing feigned performance in some populations or settings despite being a valid instrument in situations.

Investigators have known for some time that bona-fide injury and exaggeration, feigning, or malingering are not always mutually exclusive. It is particularly difficult to parcel out the feigners/exaggerators with mild head injury from those with mild head injury alone. Simulation studies on individuals with documented head-injury who are asked to exaggerate symptoms may help elucidate this area. To do this, Rogers (1997) advocated combining known-groups and simulation designs. The simulation design increases experimental rigor and allows for within-subject analyses, while known-groups design improves clinical relevance. This combined design also allows for the known-groups analysis to act as a cross-validation of the simulation portion.

Statistical Considerations

Meta-analysis of feigning detection strategies is recommended. Meta-analysis involves the statistical measure of effect sizes among previously published studies and provides a simple metric that can help clinicians decide which feigning detection may be best. Recent meta-analytic studies have helped evaluate the degree to which personal injury clients differ on neuropsychological tests from other patients not in litigation (Binder & Rohling, 1996). However, malingering research is replete with different tests, methods, techniques, and samples. These issues limit the range and usefulness of meta-analysis. As the body of literature continues to grow, future studies may be better able to examine highly similar studies, thereby reducing the number of sources of error that can confound meta-analytic data (Glass, McGaw, & Smith, 1981). Researchers and

practitioners would benefit from meta-analyses that help determine which strategies are most effective and under what conditions.

Future Research on the TOCA

The TOCA is intended to assess both feigning and ability level. Though the present study evaluated feigning detection only, the TOCA has particular promise if it can be established as a measure of both ability and feigning. Thus, the extent to which the TOCA assesses working memory (Section 1), visuo-spatial ability (Section 2), and verbal comprehension (Section 3) needs to be evaluated. Convergent validity studies with common measures of these constructs are indicated. Factor analytic studies will also be needed to help reveal the underlying constructs of each section. Underutilized items also may be revealed and included in scales to improve classification rates.

Alternative operationalizations of the strategies used in the TOCA may yield better classifications. For example, SVT appears to be more sensitive when performance is compared to 50% chance likelihood. Thus, 2-alternative items may be interspersed randomly with the 4-choice items on the TOCA.

In terms of PC, the use of running means (see Frederick & Foster, 1991) provides more points of analysis on the curve. The current study demonstrated that difference scores between adjacent levels of difficulty on the curve can reveal feigning where other PC strategies failed to detect feigning. However, only 4 levels of difficulty were available for comparisons. Increasing the number of levels of difficulty would increase the number of possible comparisons along the curve and make the strategy more sensitive to atypical changes in performance.

Expert ratings of the magnitude of errors may be helpful in establishing which incorrect items are most likely to be chosen. Martin et al. (1998) asked advanced graduate students to rate the likelihoods for each wrong response in their adaptation of the MoE strategy. A similar approach may help identify the incorrect responses on the TOCA that are more likely than others to be chosen. The preliminary data from this study suggest that this operationalization (i.e., 7.1%-MoE scales) may work particularly well.

Another promising aspect of the PC strategy may come from the fact that simulators and patients differ most at the beginning of the curve, where items are easiest. This finding is consistent with the theory that simulators are differentially likely to miss these items. This derivation of the Floor Effect, termed the "point-of-entry" by Frederick and Foster (1991) may be sensitive to feigned performance on the TOCA. Quantifiable indices of the differences at this point may prove useful.

Conclusions

In conclusion, the amount of neuropsychological research on effort, motivation, and feigning has increased in recent years, with a primary goal being to improve reliability and validity. Classification of feigning versus bona-fide injury continues to be complicated by other clinical presentations that involve feigning, poor effort, or exaggeration (e.g., Factitious Disorder, fatigue, Somatization). For example, many tests employ one cutting score that is said to differentiate feigning from true impairment. However, these scores are typically limited to one sample's performance and do not generalize to other clinical samples. The variability in presentations that resemble malingering seriously reduces the effectiveness of the score.

Nonetheless, major advances in the development of strategic detection techniques have been made. For example, many clinicians have capitalized on the efficiency of the Floor Effect, particularly on multiple-choice measures. In addition, other areas of research have made contributions to the neurocognitive feigning literature. For example, preliminary data from EEG (i.e., the frequency and amplitude of P300 waves) have shown promise in feigning detecting (Rosenfeld et al., 1999). Continued study of the utility of combining physiological and neurocognitive tests of feigning appears warranted.

The current study was a preliminary validation of the TOCA, a multi-scale measure of feigned neurocognitive impairment. Strategies were developed from information and recommendations from past studies and from new operationalizations used in the current study. The results indicate that the scales ranged in their individual utility and that combining scales improved clinical utility. In general, the data support the continued investigation of all the strategies and their clinical applications. More data regarding the clinical utility of the TOCA is needed, particularly in mild brain-injured and psychiatric populations. Though substantial gains have been made, neurocognitive feigning detection continues to be flawed by both methodological and practical limitations. Continued theory-driven research is warranted.

APPENDICES

APPENDIX A

Consent Form for Clinical Comparison Sample

Getting Motivated and Staying Motivated

Some clients referred for psychological or neuropsychological evaluations have problems staying with tasks and getting things done. Some clients get off to a great start and then lose interest. More research is urgently needed on what clinical features allow some clients able to keep working at challenging tasks and put forth their best effort.

I understand that this study will involve my completing a computer-based measure that will evaluate my efforts and abilities. This measure will take about 30 minutes to complete and is multiple-choice format. As an experimental measure, these results will have no effect on my current evaluation at Plano Rehabilitation Hospital. I understand that there are no anticipated risks from participating in this study.

I understand that the results from this measure will be compared to the other testing I am completing and to past clinical records.

I understand that my participation in the study is entirely voluntary. There is no penalty for not participating. In addition, I may withdraw from the study at any time and for any reason. Again, there will be no penalty for withdrawing. If I have any questions about the study, I may contact Dr. Jay Duhon at (972) 612-9000. I may also contact Dr. Richard Rogers or Dr. Kenneth Sewell, the principal investigators, at (940) 565-2671. This study has been reviewed and approved by the UNT Committee for the Protection of Human Subjects (940) 565-3940.

I understand that my results will be kept confidential. Any research publication will only address group results and will never identify any individual participant.

I understand and agree to the above procedures.

signed _____ date _____

witnessed _____ date _____

APPENDIX B

Screening Measure for Simulators and Controls

Before we begin, please answer the following questions. All information is confidential and will be used only for descriptive purposes:

1. Have you ever experienced a head injury in which you lost consciousness for a significant period of time? _____ If so, how long were you unconscious? _____ Did you receive medical attention? _____ Were you hospitalized? _____ If so, for how long? _____

2. What is your cumulative GPA? _____

3. What year in school are you?

   Freshman ____ Sophomore ___ Junior _____ Senior _____

4. What is your gender? _____

5. What is your age? _____

6. What hand do you use to write? _____

7. What ethnic category best describes you? African-American _____ Caucasian ___ Asian-American _____ Hispanic ____ American-Indian ____ Other _____

8. Have you ever been diagnosed with a learning disability, ADHD, or any other psychological disorder (e.g., anxiety, depression)?

APPENDIX C

Informed Consent for Simulators and Controls

Consent Form

Getting Motivated and Staying Motivated

Some people have problems staying with tasks and getting things done. Some get off to a great start and then lose interest. More research is urgently needed on what features allow some people to be able to keep working at challenging tasks and put forth their best effort.

I understand that this study will involve my completing a computer-based measure that will evaluate my efforts and abilities. This measure will take about 30 minutes to complete and is multiple-choice format. I understand that there are no anticipated risks from participating in this study.

I understand that my participation in the study is entirely voluntary. There is no penalty for not participating. In addition, I may withdraw from the study at any time and for any reason. Again, there will be no penalty for withdrawing. If I have any questions about the study, I may contact Scott Bender, at (940) 387-2291. Or, I may reach Dr. Richard Rogers or Dr. Kenneth Sewell, at (940) 565-2671. This study has been reviewed and approved by the UNT Committee for the Protection of Human Subjects (940) 565-3940.

I understand that my results will be kept confidential. Any research publication will only address group results and will never identify any individual participant.

I understand and agree to the above procedures.

signed _____ date _____

witnessed _____ date _____

APPENDIX D

Please answer the following brief questions regarding the test:

1. What were you asked to do? _____

_____

_____

2. Did the instructions make sense to you? Were they in any way confusing? _____

_____

_____

3. Were you asked to fake an injury? _____ If not, please go to question #4.

3a. If so, what percentage of the time did you *carefully* follow the instructions?

_____%. How careful were you to follow the instructions?

Check one: Not at all _____ Somewhat _____ Quite _____ Very _____

3b. How successful at fooling the test were you? Very successful _____ Somewhat successful _____ A little successful _____ Not very successful _____

3c. If you feel you were not able to fake convincingly, what hindered you?

I am too honest _____ I didn't understand the instructions _____ Too hard to fake on this test _____ Too easy _____

Other _____ (explain) _____

3d. If you feel you were successful, what helped you fake? *Check all that apply:*

Knowledge of brain function _____ I know people with brain damage _____

Other _____ (explain) _____

3e. In faking the test, what did you do to appear brain-injured?

Answered incorrectly on purpose _____

Slowed down my responses _____

Tried to appear memory impaired _____

Tried to appear inattentive _____

4. At what level would you rate your knowledge of brain function?

Highly specialized _____ good _____ average _____ poor _____

5. What did you do before the test, during the preparation time? _____

_____

6. How would you rate your efforts?

Great effort _____ Half an effort _____ Little effort _____ No effort _____

7. Did you try any harder because there was an incentive? _____ *Check one:*

Not at all_____ A little _____ Somewhat harder _____ Quite a bit harder _____

Much harder _____

APPENDIX E

TOCA

Test of Cognitive Abilities

This is a test designed to measure how you think and solve problems. Please follow these simple rules:

1. Don't rush yourself; you make take as much time as you need.
2. Don't get "stuck;" if you can't figure it out, choose the best answer you can.
3. Don't leave blanks; you can get partial credit even if you don't have the right answer.

Some items are easy; some items are difficult; and some items are almost impossible. Don't worry about getting them all right. Just do your best job.

To proceed, press the spacebar.

APPENDIX F

Instructions for Normal Control Group

You are taking a test that measures problem-solving skills and attention. Please answer all questions to the best of your ability. Simply follow the directions provided by the examiner and the computer.

- There is a reward of $20 for each of the three highest scores! Do your best to win.
- You will be asked to fill out a brief questionnaire at the end of the test.
- The entire time should take approximately 45 minutes.
- Be sure to collect proof of extra credit before leaving.

Thank you for your participation!

APPENDIX G

TOCA Intake Form

Name: _____ Subject #: _____ Date of Testing: _____

DOB: _____ Gender: _____ (1 = female; 2 = male)

Age: _____ Date of Injury: _____

Ethnicity: _____ (1=African Amer; 2=Hispanic; 3=Caucasian; 4=Asian; 5=Amer Indian; 6=other)

Education level: _____ (0 = no high school; 1 = some HS; 2 = HS diploma; 3 = some college; 4 = bachelor degree; 5 = graduate school)

Employment: _____ (0 = currently unemployed; currently part-time = 1; currently full-time = 2; student = 3)

Handedness: _____ (1 = right; 2 = left) Litigating: _____ (1 = yes; 2 = no)

Diagnosis at Intake: Medical _____ Psych Axis I: _____

_____ _____

_____ Axis II: _____

Location of injury: _____

Status: _____ (1 = inpatient; 2 = outpatient)

Pre-existing conditions: _____

Notes: _____

_____

_____

APPENDIX H

Debriefing Interview for Clinical Comparison Sample

1. How would you rate your motivation during the test?

Very Poor ___ Poor ___ Average ___ Pretty Good _ Excellent ___

2. What do you think caused you to be motivated or not?

3. Did you understand what you were to do?

4. Was any part of the testing confusing?

5. Do you have any questions?

APPENDIX I

Responses to Open-ended Question in Debriefing, "If you feel you were successful, what helped you fake?"

"I considered it a challenge to fake convincingly." (x 3)

"Letting my mind go blank" (x 2)

"It was the lecture in one of my classes this week."

"Focusing on the advice in the scenario"

"I though it was a for a combination of reasons." (reasons not listed)

"I thought about how to fool the test"

"Pretending to have brain damage"

"I pretended I was an actor"

"Fidgeting a lot"

"Asking a lot of unnecessary questions"

APPENDIX J

Group Mean and Effect Size (Cohen's d) Comparison with F statistic for the RT, Section Score, and RT x Section Score Scales

	Group					
	SIM		Honest			
Strategy	CS	NCS	CL	NC	F	Cohen's d
Section 1						
RT	21.38_{a} (7.17)	21.63_{a} (5.46)	40.03_{b} (19.32)	26.21_{a} (6.27)	9.54^{e}	1.42
Section Score	155.76_{a} (44.30)	155.17_{a} (45.41)	176.29_{a} (56.94)	205.64_{b} (27.81)	6.13^{d}	0.41
RT X Section Score	3008.02_{a} (1332.10)	3481.68_{ab} (1528.36)	7114.71_{b} (3866.09)	5363.85_{b} (1422.31)	16.79^{e}	1.46
Section 2						
RT	14.43_{a} (4.73)	13.95_{a} (4.78)	19.06_{b} (4.91)	16.46_{ab} (3.91)	4.26^{d}	1.00
Section Score	33.76_{a} (10.34)	32.21_{a} (9.54)	27.29_{a} (10.05)	35.64_{b} (12.45)	3.44^{d}	0.57
RT X Section Score	518.69_{a} (282.24)	464.55_{a} (264.17)	537.60_{a} (241.20)	618.94_{a} (329.77)	1.35	0.18
Section 3						
RT	12.76_{ab} (4.25)	12.07_{a} (3.56)	16.29_{b} (10.22)	9.18_{a} (3.22)	3.38^{d}	0.55
Section Score	82.12_{a} (23.93)	82.69_{a} (24.67)	103.09_{b} (16.07)	109.64_{b} (7.56)	13.27^{e}	1.02
RT X	1165.61_{a}	982.98_{a}	1756.31_{b}	999.63_{a}	12.51^{e}	1.05

Section Score	(591.77)	(353.10)	(830.18)	(339.39)		
3 Sections Combined						
RT	47.23_{ab} (13.05)	47.65_{b} (10.41)	79.18_{a} (27.50)	51.85_{b} (9.32)	8.80_{e}	1.58
Section Score	271.63_{a} (66.91)	270.07_{a} (68.54)	302.76_{a} (72.44)	350.91_{b} (38.64)	8.70_{d}	0.45
RT x Section Score	12957.81_{a} (4609.33)	13209.91_{a} (5074.14)	24039.48_{b} (8010.47)	18182.37_{c} (3939.81)	24.41_{e}	1.68

Notes. Means with different subscripts are significantly different by Tukey comparison, $p < .05$. Cohen's d was calculated for CL versus all simulators (CS + NCS = SIM).
For Groups, CL = Clinical, CS = Cautioned Simulators, NCS = Non-cautioned Simulators, and NC = Normal Controls.
$_{d}$ For F ratios, $p < .01$
$_{e}$ For F ratios, $p < .001$

APPENDIX K

A trend for Caucasian Americans to score slightly higher than Hispanic Americans on the TOCA was noted, though this difference was not statistically significant. Given this trend, whether differences emerged on the detection scales themselves was also examined. Among normal control (NC) participants, ethnicity did not significantly affect performance on the most effective detection scales (see Table below). Given the large difference in sample sizes (i.e., n = 3 Hispanic Americans vs. n = 19 Caucasian Americans), these data are considered exploratory and should be interpreted cautiously.

Table 42

Means (and SDs) for Hispanic American and Caucasian American Normal Controls on RT Total, 7.1%-MoE3, FE-95%-I, and Rate of Decay

	Ethnicity			
Scale	HA	CA	t	p
RT Total	56.45 (9.72)	41.36 (13.12)	.55	.59
7.1%-MoE3	0.00 (0.00)	.42 (.84)	.85	.40
FE-95%-I	7.00 (0.00)	6.84 (.37)	-.72	.48
Rate of Decay	145.33 (40.0)	121.05 (32.36)	-.98	.25
n	3	19		

Note. HA = Hispanic American and CA = Caucasian American

In contrast to the findings in NC, a significant difference between Caucasian American and Hispanic American performance was noted among simulators (see Table below). Hispanic American simulators performed more poorly on the FE-95%-I detection scale than Caucasian Americans (i.e., they produced scale elevation indicative feigning more often), $\underline{p} = .01$. Specifically, 80.0% of Hispanic American simulators scored below the recommended cutting score on the scale, while 28.2% of Caucasian Americans fell below the cutting score. No differences were found on the other most effective detection scales. It appears that Hispanic Americans missed easy items more often than Caucasian Americans in order to appear impaired. This finding may suggest that Caucasian American and Hispanic Americans engage in different techniques when attempting to portray themselves as impaired. Alternatively, they may differ in sophistication when feigning impairment. Research has yet to explore the implications of these potential differences for scale development and clinical practice.

APPENDIX L

A trend was noted for participants with a history of mental disorder to score slightly lower on the TOCA than those without such a history, p = .10. Whether these participants were more vulnerable to detection as a function of their history of mental disorder was examined. As can be seen in the table below, the groups did not differ in performance on the most effective detection scales. In addition, the proportion of participants with a history of mental disorder who scored below the recommended cutting score was not significantly different from the proportion of those without a history of mental disorder who fell below this score (Contingency Coefficient = .253, p = .84).

Table 43

Means (and SDs) for Those With a History of Mental Disorder Versus Those Without on RT Total, 7.1%-MoE3, FE-95%-I, and Rate of Decay

Scale	MD	NMD	t	p
RT Total	51.57 (10.20)	47.12 (13.59)	1.01	.32
7.1%-MoE3	19.27 (8.87)	16.08 (12.21)	.81	.42
FE-95%-I	5.64 (1.29)	5.23 (1.87)	.68	.49
Rate of Decay	84.18 (39.88)	83.05 (51.99)	.07	.95

Note. MD = History of Mental Disorder and NMD = No History of Mental Disorder.

REFERENCES

Allen, L. M., Conder, R., Green, P., & Cox, D. R. (1997). Computerized Assessment of Response Bias. Durham, NC: Cognisyst, Inc.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed., text revision). Washington, DC: Author.

Amin, K., & Prigatano, G. P. (1993). Digit Memory Test: Unequivocal cerebral dysfunction and suspected malingering. Journal of Clinical & Experimental Neuropsychology. Vol 15(4), 537-546.

Anastasi, A. & Urbina, U. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, W., Trethowan, W., & Kenna, J. (1959). An experimental investigation of simulation and pseudo-dementia. Acta Psychiatrica et Neurologica Scandinavica, 34(132; whole issue).

Arnett, P. A., Hammeke, T. A., & Schwartz, L. (1995). Quantitative and qualitative performance on Rey's 15-Item Test in neurological patients and dissimulators. The Clinical Neuropsychologist, 9(1), 17-26.

Bash, I. Y., & Albert, M. (1980). The determination of malingering. Annals of New York Academy of Sciences, 347, 86-99.

Beetar, J. T., & Williams, J. M. (1995). Malingering response styles on the Memory Assessment Scales and symptom validity tests. Archives of Clinical Neuropsychology, 10(1), 57-72.

Bernard, L. C. (1990). Prospects for faking believable memory deficits on neuropsychological tests and the use of incentives in simulation research. Journal of Clinical and Experimental Neuropsychology, 12, 715-728.

Bernard, L. C., Houston, W., & Natoli, L. (1993). Malingering on neuropsychological tests: Potential objective measures. Journal of Clinical Psychology, 49, 45-53.

Bernard, L. C., McGrath, M. J., & Houston, W. (1996). The differential effects of simulating malingering, closed head injury, and other CNS pathology on the Wisconsin

Card Sorting Test: Support for the "Pattern of Performance" hypothesis. Archives of Clinical Neuropsychology, 11(3), 231-245.

Berg, E. (1948). A simple objective treatment for measuring flexibility in thinking. Journal of General Psychology, 39, 15-22.

Binder, L. M. (1992b). Malingering detected by forced choice testing of memory and tactile sensation: A case report. Archives of Neuropsychology, 7, 155-163.

Binder, L. M. (1993). Assessment of malingering after mild head trauma with the Portland Digit Recognition Test. Journal of Clinical and Experimental Neuropsychology, 15, 170-182.

Binder, L. M. (1995; November). Assessment of Functional Problems. Paper presented at the meeting of the National Academy of Neuropsychology, San Antonio, TX.

Binder, L. M., & Rohling, M. (1996). Money matters: A meta-analytic review of the effects of financial incentives on recovery after closed-head injury. American Journal of Psychiatry, 153, 7-10.

Binder, L. M., & Willis, S. C. (1991). Assessment of motivation after financially compensable minor head trauma. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3, 175-181.

Binks, P. G., Gouvier, W. D., & Waters, W. F. (1997). Malingering detection with the Dot Counting Test. Archives of Clinical Neuropsychology, 12(1), 41-46.

Boone, K. B., Lu, P., Sherman, D., Palmer, B., Back, C., Shamieh, E., Warner-Chacon, K., & Berman, N. G. (2000). Validation of a new technique to detect malingering of cognitive symptoms: The b Test. Archives of Clinical Neuropsychology, 15(3), 227-241.

Brady, J., & Lind, D. (1961). Experimental analysis of hysterical blindness. Archives of General Psychiatry, 4, 331-339.

Brown, L., Sherbenou, R. J., & Johnson, S. K. (1982). Test of Nonverbal Intelligence: A language-free measure of cognitive ability. Austin, TX: Pro-Ed.

Cercy, S. P., Schretlen, D. J., & Brandt, J. (1997). Simulated amnesia and the pseudo-memory phenomenon. In R. Rogers (Ed.), Clinical assessment of malingering and deception (2nd ed., pp. 85-107). New York: The Guilford Press.

Coleman, R. D., Rapport, L. J., Millis, S. R., Ricker, J. H., & Farchione, T. J. (1998). Effects of Coaching on detection of malingering on the California Verbal Learning Test. Journal of Clinical and Experimental Neuropsychology, 20(2), 201-210.

Cullum, C., Heaton, R. K., & Grant, I. (1991). Psychogenic factors influencing neuropsychological performance: Somatoform disorders, factitious disorders, and malingering. In H. Doerr & A. Carlin (Eds.), Forensic neuropsychology: Legal and scientific bases (pp. 79-89). New York: The Guilford Press.

Cunnien, A. J. (1997). Psychiatric and medical syndromes associated with deception. In R. Rogers (Ed.), Clinical assessment of malingering and deception (2nd ed., pp. 23-46). New York: The Guilford Press.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. (1987). California Verbal Learning Test - Adult version. San Antonio, TX: Psychological Corporation.

Etcoff, L. M., & Kampfer, K. M. (1996). Practical guidelines in the use of symptom validity and other psychological tests to measure malingering and symptom exaggeration in traumatic brain injury cases. Neuropsychology Review, 6(4), 171-201.

Franzen, M. D., Iverson, G. L., & McCracken, L. M. (1990). The detection of malingering in neuropsychological assessment. Neuropsychological Review, 1(3), 247-279.

Franzen, M. D., & Martin, N. (1996). Do people with knowledge fake better? Applied Neuropsychology, 3, 82-85.

Frederick, R. I., Carter, M., & Powel, J. (1995). Adapting symptom validity testing to evaluate suspicious complaints of amnesia in medicolegal evaluations. Bulletin of the American Academy of Psychiatry and the Law, 23, 231 – 237.

Frederick, R. I., & Crosby, R. D. (2000). Development and validation of the Validity Indicator Profile. Law and Human Behavior, 24,(1), 59 – 82.

Frederick, R. I., & Foster, H. G. (1991). Multiple measures of malingering on a forced choice test of cognitive ability. Psychological Assessment, 3(4), 596-602.

Frederick, R. I., Sarfaty, S. D., Johnston, D., & Powel, J. (1994). Validation of a detector of response bias on a forced-choice test of nonverbal ability. Neuropsychology, 8(1), 118-125.

Frederick, R. I., & Foster, H. G. (1997). The Validity Indicator Profile. Minneapolis: National Computer Systems.

Glass, G., McGaw, B., & Smith, M. (1981). Meta-analysis in social research. Beverly Hills: Sage Publications.

Goldberg, J. O, & Miller, A. R. (1986). Performance of psychiatric inpatients and intellectual deficient individuals on a task that assesses the validity of memory complaints. Journal of Clinical Psychology, 42(5), 792-795.

Golden, C. J., Hammeke, T. A., & Purisch, A. D. (1980). The Luria-Nebraska Neuropsychological Test Battery: Manual. Los Angeles: Western Psychological Services.

Gouvier, W. D. (1999). Base rates and clinical decision making in neuropsychology. In J. Sweet (Ed.), Forensic neuropsychology: Fundamentals and practice (pp. 27-37). Lisse: Swets and Zeitlinger.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. Psychological Assessment, 6, 218-224.

Griffin, G. A., Glassmire, D. M., Henderson, E. A., & McCann, C. (1997). Rey II: Redesigning the Rey screening test of malingering. Journal of Clinical Psychology, 53(7), 757-766.

Griffin, G. A., Normington, J., May, R., & Glassmire, D. M. (1996). Assessing dissimulation among Social Security income claimants. Journal of Consulting and Clinical Psychology, 64, 1425-1430.

Gudjonsson, G. H., & Shackleton, H. (1986). The pattern of scores on Raven's Matrices during 'faking bad' and 'non-faking' performance. British Journal of Clinical Psychology, 25 35-41.

Guilmette, T. J., Hart, K. J., & Giuliano, A. J. (1993). Malingering detection: The use of a forced-choice method in identifying organic versus simulated memory impairment. The Clinical Neuropsychologist, 7(1), 59-69.

Guilmette, T. J., Hart, K. J., Giuliano, A. J., & Leininger, B. (1994). Detecting simulated memory impairment: Comparison of the Rey Fifteen-Item Test and the Hiscock Forced-Choice Procedure. The Clinical Neuropsychologist, 8(3), 283-294.

Hair, J. F., Anderson, R, Tatham, R., & Black, W. C. (1995). Multivariate Data analysis (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Heaton, R. K., Smith, H. H., Lehman, R. A., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. Journal of Consulting and Clinical Psychology, 46, 892-900.

Hiscock, C. K., & Hiscock, M. (1989). Refining the forced-choice method for the detection of malingering. Journal of Clinical and Experimental Neuropsychology, 11, 967-974.

Hiscock, C. K., Branham, J., & Hiscock, M. (1994). Detection of feigned neurocognitive impairment: The two-alternative forced-choice method compared with selected conventional tests. Journal of Psychopathology and Behavioral Assessment, 16(2), 95-109.

Iverson, G. L., & Franzen, M. D. (1996). Using multiple objective memory procedures to detect simulated malingering. Journal of Clinical and Experimental Neuropsychology, 18, 38-51.

Iverson, G. L., Franzen, M. D., & McCracken, L. M. (1991). Evaluation of an objective assessment technique for the detection of malingered memory deficits. Law and Human Behavior, 15, 667-676.

Johnson, J. L., & Lesniak-Karpiak, K. (1997). The effect of warning on malingering on memory and motor tasks in college samples. Archives of Clinical Neuropsychology, 12(3), 231-238.

Jonas, J., & Pope, H. (1985). The dissimulating disorders: A single diagnostic entity? Comprehensive Psychiatry, 26, 58-62.

Kazdin, A. (1992). Research design in clinical psychology (2nd ed.). Boston: Allyn & Bacon.

Kolb, B., & Wishaw, I. Q. (1996). Fundamentals of human neuropsychology (4th ed.). New York: WH Freeman and Company.

Lees-Haley, P. R., & Brown, R. (1993). Neuropsychological complaint base rates of 170 personal injury claimants. Archives of Clinical Neuropsychology, 8, 203-209.

Lees-Haley, P. R., Smith, H., Williams, C., & Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical survey. Archives of Clinical Neuropsychology, 11(1), 45-51.

Lezak, M. D. (1995). Neuropsychological assessment (3rd ed.). New York: Oxford.

McCaffery, R., Williams, A., Fisher, J., & Laing, L. (1997). The practice of forensic neuropsychology: Meeting challenges in the courtroom. New York: Plenum Press.

McKinzey, R. M., Podd, M. H., Krehbiel, M., Mensch, A., & Trombka, C. (1997). Detection of malingering on the Luria-Nebraska Neuropsychological Battery: An initial and cross-validation. Archives of Clinical Neuropsychology, 12(5), 505-512.

Martell, D. (1992). Forensic neuropsychology and the criminal law. Law and Human Behavior, 16(3), 313-336.

Martin, R. C., Bolter, J. F., Todd, M. E., Gouvier, W. D., & Niccolls, R. (1993). Effects of sophistication and motivation on the detection of malingered memory performance using a computerized forced-choice task. Journal of Clinical and Experimental Neuropsychology, 15(6), 867-880.

Martin, R. C., Franzen, M. D., & Orey, S. (1998). Magnitude of error as a strategy to detect feigned memory impairment. The Clinical Neuropsychologist, 12(1), 84-91.

Maxmen, J., & Ward, N. (1995). Factitious Disorders. In Essential psychopathology and its treatment (2nd ed., pp. 304-309). New York: WW Norton.

Mensch, A. J., & Woods, D. J. (1986). Patterns of feigning brain damage on the LNNB. International Journal of Clinical Neuropsychology, 8, 59-63.

Meyers, J., & Volbrecht, M. (1998). Validation of reliable digits for detection of malingering. Assessment, 5, 303-307.

Miller, H., & Cartlidge, N. (1972). Simulation and malingering after injuries to the brain and spinal cord. Lancet, 1, 580-586.

Millis, S. R. (1992). The Recognition Memory Test in the detection of malingered and exaggerated memory deficits. Clinical Neuropsychologist, 6, 405-413.

Millis, S. R., & Kler, S. (1995). Limitations of the Rey fifteen-item test in the detection of malingering. The Clinical Neuropsychologist, 9, 241-244.

Millis, S. R., Putnam, S. H., Adams, K. H., & Ricker, J. H. (1995). The California Verbal Learning Test in the detection of incomplete effort in neuropsychological evaluation. Psychological Assessment, 7, 463-471.

Mittenberg, W., Azrin, R., Millsaps, C., & Heilbronner, R. (1993). Identification of malingered head injury on the Wechsler Memory Scale - Revised. Psychological Assessment, 5, 34-40.

Mittenberg, W., Theroux-Fichera, S., Zielinski, R., & Heilbronner, R. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale-revised. Professional Psychology: Research and Practice, 26(5), 491-498.

Nies, K. J., & Sweet, J. J. (1994). Neuropsychological assessment and malingering: A critical review of past and present strategies. Archives of Clinical Neuropsychology, 9(6), 501-552.

Pachana, N., Boone, K., & Ganzell, S. (1998). False positive errors on selected tests of malingering. American Journal of Forensic Psychology, 16(2), 17-25.

Pankratz, L. (1979). Symptom validity testing and symptom retraining: Procedures for the assessment and treatment of functional sensory deficits. Journal of Consulting and Clinical Psychology, 47, 409-410.

Pankratz, L., & Binder, L. M. (1997). Malingering on intellectual and neuropsychological measures. In Rogers, R. (Ed.), Clinical assessment of malingering and deception (pp. 223-236). New York: Guilford.

Posford, J., Willmott, C., Rothwell, A., Cameron, P., Helly, A., Nelms, R., Curran, C., & Ng, K. (2000). Factors influencing outcome following mild traumatic brain injury in adults. Journal of the International Neuropsychological Society, 6, 568-579.

Raven, J. (1958). Guide to the Raven's Matrices test. London: H. Lewis & Co.

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments with the Test of Memory Malingering. Psychological Assessment, 10(1), 10-20.

Reitan, R. M., & Wolfson, D. (1993). The Halstead-Reitan Neuropsychological Battery: Theory and clinical interpretation (2nd ed.). Tuscon, AZ: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1996). The question of validity of neuropsychological test scores among head-injured litigants: Development of a dissimulation index. Archives of Clinical Neuropsychology, 11(7), 573-580.

Reitan, R. M., & Wolfson, D. (1997). Consistency of neuropsychological test scores of head-injured subjects involved in litigation compared with head-injured subjects not involved in litigation: Development of the test-retest index. The Clinical Neuropsychologist, 11(1), 69-76.

Rey, A. (1964). L'Examen clinique en psychologie [The clinical exam in psychology]. Paris: Presses Universitaires de France.

Rogers, R. (1990a). Development of a new classificatory model of malingering. Bulletin of the American Academy of Psychiatry and Law, 18, 323-333.

Rogers, R. (1990b). Models of feigned mental illness. Professional Psychology: Research and Practice, 21, 182-188.

Rogers, R. (1997). Researching dissimulation. In Rogers, R. (Ed.), Clinical assessment of malingering and deception (2nd ed., pp. 398-426). New York: Guilford.

Rogers, R. (1996). The Test of Cognitive Abilities.©

Rogers, R. Bagby, R. M., & Rector, N. (1989). Diagnostic legitimacy of factitious disorder with psychological symptoms. American Journal of Psychiatry, 146, 1312-1314.

Rogers, R., Bagby, R. M., & Vincent, A. (1994). Factitious disorders with predominantly psychological signs and symptoms: A conundrum for forensic experts. Journal of Psychiatry and Law, 22, 91-106.

Rogers, R., & Cruise, K. R. (1998). Assessment of malingering with simulation designs: Threats to external validity. Law & Human Behavior. Vol 22(3), 273-285.

Rogers, R., Harrell, E. H., & Liff, C. D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. Clinical Psychology Review, 13, 255-274.

Rogers, R., Salekin, R. T., Sewell, K. W., Goldstein, A., & Leonard, K. (1998). A comparison of forensic and nonforensic malingerers: A prototypical analysis of explanatory models. Law & Human Behavior. Vol 22(4), 353-367.

Rogers, R., & Reinhardt, V. R. (1998). Conceptualization and assessment of secondary gain. In G. Koocher & J. Norcross (Eds.), Psychologist's desk reference (pp. 57-62) New York: Oxford.

Rogers, R., Sewell, K. W., & Goldstein, A. (1994). Explanatory models of malingering: A prototypical analysis. Law and Human Behavior, 18, 543-552.

Rogers, R., & Vitacco, M. (in press). Forensic assessment of malingering and related response styles.

Rose, F. E., Hall, S., & Szalda-Petree, A. D. (1995). Portland Digit Recognition Test-computerized: Measuring response latency improves the detection of malingering. The Clinical Neuropsychologist, 9, 124-134.

Rose, F. E., Hall, S., Szalda-Petree, A. D., & Bach, P. (1998). A comparison of four tests of malingering and the effects of coaching. Archives of Neuropsychology, 13(4), 349-363.

Rosenfeld, B., Sands, S. A., & Van Gorp, V. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. Archives of Clinical Neuropsychology, 15(4), 349-359.

Rosenfeld, J. P., Ellwanger, J. W., Nolan, K., Wu, S., Bermann, R. G., & Sweet, J. J. (1999). P300 scalp amplitude distribution as an index of deception in a simulated cognitive deficit model. International Journal of Psychophysiology, 33(1), 3-19.

Ruff, R. M., Camenzuli, L, & Mueller, J. (1996). Miserable minority: emotional risk factors that influence the outcome of a mild traumatic brain injury. Brain Injury, 10, 551-566.

Schacter, D. L. (1986). On the relation between genuine and simulated amnesia. Behavioral Sciences and the Law, 4, 47-64.

Schretlen, D. J. (1988). The use of psychological tests to identify malingered symptoms of mental disorder. Clinical Psychology Review, 8, 451-476.

Schretlen, D. J., & Arkowitz, D. (1990). A psychological test battery to detect prison inmates who fake insanity or mental retardation. Behavioral Sciences and the Law, 8(1), 75-84.

Schretlen, D. J., Brandt, J., Krafft, L., & Van Gorp, W. G. (1991). Some caveats in using the Rey 15-Item Memory Test to detect malingered amnesia. Psychological Assessment, 3(4), 667-672.

Shipley, W. (1946). The Institute of Living Scale. Los Angeles: Western Psychological Services.

Slick, D. J., Hopp, G., Strauss, E., Hunter, M., & Pinch, D. (1994). Detecting dissimulation: Profiles of simulated malingerers, traumatic brain-injured individuals, and normal controls on a revised version of Hiscock and Hiscock's forced-choice memory test. Journal of Clinical and Experimental Neuropsychology, 16(3), 472-481.

Slick, D. J., Hopp, G., Strauss, E., & Spellacy, F. (1996). Victoria Symptom Validity Test: Efficiency for detecting feigned memory impairment and the relationship to neuropsychological tests and MMPI-2 validity scales. Journal of Clinical and Experimental Neuropsychology, 18(6), 911-922.

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. The Clinical Neuropsychologist, 13(4), 545-561.

Smith, G. (1992). Detection of malingering: A validation study of the SLAM test. Unpublished doctoral dissertation, University of Missouri-St. Louis.

Sutherland, A., & Rodin, G. (1990). Factitious disorders in a general hospital setting: Clinical features and a review of the literature. Psychosomatics, 31, 392-399.

Sweet, J. J. (1999). Malingering: Differential diagnosis. In J. J. Sweet (Ed.), Forensic Neuropsychology (pp. 255-286). Lisse: Swets and Zeitlinger.

Sweet, J. J., Wolfe, P., Sattlberger, E., Numan, B., Rosenfeld, J. P., Clingerman, S., & Nies, K. J. (2000). Further investigation of traumatic brain injury versus insufficient effort with the California Verbal Learning Test. Archives of Clinical Neuropsychology, 15(2), 105-113.

Tenhula, W. N., & Sweet, J. J. (1996). Double cross-validation of the Booklet Category Test in detecting malingered traumatic brain injury. The Clinical Neuropsychologist, 10, 104-116.

Tombaugh, T. N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. Psychological Assessment, 9(3), 260-268.

Trueblood, W. (1994). Qualitative and quantitative characteristics of malingered and other invalid WAIS-R and clinical memory data. Journal of Clinical and Experimental Neuropsychology, 16(4), 597-607.

Trueblood, W., & Binder, L. M. (1997). Psychologists' accuracy in identifying neuropsychological test protocols of clinical malingerers. Archives of Clinical Neuropsychology, 12(1), 13-27.

Trueblood, W., & Schmidt, M. (1993). Malingering and other validity considerations in the neuropsychological evaluation of mild head injury. Journal of Clinical and Experimental Neuropsychology, 15, 578-590.

Van Gorp, W. G., Humphrey, L. A., Kalechstein, A., Brumm, V. L., McMullen, W. J., Stoddard, M., & Pachana, N. A. (1999). How well do standard clinical neuropsychological tests identify malingering? A preliminary analysis. Journal of Clinical & Experimental Neuropsychology, 21(2), 245-250.

Warrington, E. K., (1984). Recognition Memory Test. Berkshire, UK: NFER-Nelson.

Wechsler, D. (1981). Wechsler Adult Intelligence Scale - Revised manual. New York: The Psychological Corporation.

Wechsler, D. (1987). Wechsler Memory Scale-Revised manual. San Antonio, TX: The Psychological Corporation.

Weissman, H. (1990). Distortions and deceptions in self-presentation: Effects of protracted litigation in personal injury cases. Behavioral Sciences and the Law, 8, 67-74.

Wiggins, E. C., & Brandt, J. (1988). The detection of simulated amnesia. Law and Human Behavior, 12, 57-78.

Williams, J. (1991). Memory Assessment Scales. Odessa, FL: Psychological Assessment Resources.

Wogar, M. A., Van den Broek, M. D., Bradshaw, C. M., & Szabaldi, E. (1998). A new performance-curve method for the detection of simulated cognitive impairment. British Journal of Clinical Psychology, 37(3), 327-339.

Youngjohn, J., Burrows, L., & Erdal, K. (1995). Brain damage or compensation neurosis? The controversial post-concussion syndrome. The Clinical Neuropsychologist, 9(2), 112-123.

Youngjohn, J. R., Lees-Haley, P. R., & Binder, L. M. (1999). Comment: Warning malingerers produces more sophisticated malingering. Archives of Clinical Neuropsychology, 14(6), 511 – 515.